Estimation in Slow Mixing, Long Memory Channels

Meysam Asadi, Ramezan Paravi Torghabeh and Narayana P. Santhanam

Abstract

We consider estimation of finite alphabet channels with memory where the transition probabilities (*channel parameters*) from the input to output are determined by prior outputs (*state of the channel*). While the channel is unknown, we observe the joint input/output process of the channel—we have n *i.i.d* input symbols and their corresponding outputs. Motivated by applications related to the backplane channel, we want to estimate the channel parameters as well as the stationary probabilities for each state.

Two distinct problems that complicate estimation in this setting are (i) long memory, and (ii) *slow mixing* which could happen even with only one bit of memory. In this setting, any consistent estimator can only converge pointwise over the model class. Namely, given any estimator and any sample size n, the underlying model could be such that the estimator performs poorly on a sample of size n with high probability. But can we look at a length-n sample and identify *if* an estimate is likely to be accurate?

Since the memory is unknown a-priori, a natural approach, known to be consistent, is to use a potentially coarser model with memory $k_n = \alpha_n \log n$, where α_n is a function that grows $\mathcal{O}(1)$. While effective asymptotically, the situation is quite different when we want answers with a length-n sample, rather than just consistency. Combining results on universal compression and Aldous' coupling arguments, we obtain sufficient conditions (even for slow mixing models) to identify when naive (i) estimates of the channel parameters and (ii) estimates related to the stationary probabilities of the channel states are accurate, and bound their deviations from true values.

Index Terms

Context-tree weighting, Coupling, Markov processes, Pointwise consistency, Universal compression.

I. Introduction

Science in general, information theory and learning in particular, have long concerned themselves with parsimonious descriptions of data. One of the driving motivations is that if we understand the statistical model underlying the data, we should be able to describe the data using the fewest possible bits—a concept that is formalized by the notion of entropy and that forms the foundation for quantifying information.

A more interesting situation arises when the underlying statistical model is unknown. Depending on the application at hand, we then consider a set $\mathcal{P}$ of all possible models that are consistent with what is known about the application—say all *i.i.d* or Markov models. It is often possible that we can still describe the data, without knowing the underlying model other than that it belongs to $\mathcal{P}$, using almost as succinct a description as if we knew the underlying statistics. This notion is formalized as *universal compression* [1], [2]. Of course, the universal description always incurs a penalty—the *redundancy* or excess bits beyond the minimum needed if we knew the model.

Closely related to universal compression is the concept of Minimum Description Length (MDL), where the redundancy above is interpreted as the complexity of the model class. A natural question then is—if the redundancy is small (*i.e.,* can be upper bounded uniformly for all models in $\mathcal{P}$ by a function sublinear in

The authors are with the Department of Electrical Engineering, University of Hawai'i, Mānoa.

the data length), in the process of obtaining a universal description did we somehow *estimate* the unknown model in $\mathcal{P}$?

While the general answer is no, universal compression and estimation often do go hand in hand, at least in several elementary settings. Suppose $\mathcal{P}$ is the collection of all distributions over a finite set $\mathcal{A}$, and the data at hand is a length-n sequence in $\mathcal{A}^n$ (the set of all length-n strings of symbols from $\mathcal{A}$) obtained by independent and identically distributed (*i.i.d*) sampling from an unknown distribution in $\mathcal{P}$. It is then possible to estimate the underlying distribution using a universal description, as is the case with the universal estimators by [3] or [4], [5]). As a more complex example, the Good Turing estimator (see [6]) can also be interpreted as being obtained from such a universal description [7] in a more general setting—data is exchangeable, rather than *i.i.d*.

Even with Markov processes, the connections between universal compression and estimation become more subtle. At the gross level, compression guarantees hold uniformly over Markov model classes but usually not for estimation. Despite this apparent dichotomy, our estimation results rely fundamentally on universal compression being possible. Consider a length-n sample obtained from an unknown source in the class of binary Markov sources with memory one. No matter what the source is or how the sample looks like, there are universal algorithms which describe the sample using at most $\mathcal{O}(\log n)$ bits more than the best description if the source were known. Yet, as we will see, irrespective of how large n is, we may not be in a position to reliably provide estimates of the stationary probabilities of 0s and 1s.

Let the transition probability from 1 to 0 in our memory-1 source be $\epsilon \ll 1/n$. By changing the transition probability from 0 to 1 appropriately, we can vary the stationary probabilities of 1s and 0s in a wide range, without changing how a length-n sample will look like.

Example 3 in Section **??** gives two such binary, one-bit memory Markov sources with stationary probabilities (1/2,1/2) and (2/3,1/3) respectively. But, if we start from 1, *both* sources will, with high probability, yield a sequence of n 1s. We cannot distinguish between these two sources with a sample of this size—and hence it is futile to estimate stationary probabilities from this sample. This particular phenomenon, where the number of times each state (0 and 1 here) appears is very different from their stationary probabilities is often formalized as *slow mixing*, see [8].

In this paper, we deal with estimation and operation of channels where, both the input and output are from a finite alphabet. The channel state depends on prior outputs alone and there is no feedback. We are motivated by the *backplane channel*, which we will describe shortly. We ask—given n inputs and their corresponding outputs, can we estimate the transition probabilities from the input to output, as well as the stationary probabilities of various states (or sets of states)?

We emphasize at the outset that we do *not* exclude slow mixing of the channel evolution. Instead, our philosophy will be: given n samples, what is the best answer we can give, if anything? As the above example shows, we have an estimation problem where any estimator can only converge pointwise to the true values, rather than uniformly over the model class. One way to get around this impasse is to add restrictions on the model space as is done in most prior work. However, very few such restrictions are justified in our application. So we take a different approach: can look at some characteristics of our length-n sample and say if any estimates are doing well?

Say, for the sake of a concrete example, that we have a sample $\mathbf{x}_1$ from a binary Markov process with memory 1, with $n - \log n$ 1s followed by a string of $\log n$ 0s—so perhaps, this may have come from a source as in Example 3. As we saw, it is futile to estimate stationary probabilities in this case. Contrast this sample with a new sample $\mathbf{x}_2$, also with $n - \log n$ 1s and $\log n$ 0s, but $\mathbf{x}_2$ has 0s spread uniformly in the sequence. Unlike in the case of $\mathbf{x}_1$, upon seeing $\mathbf{x}_2$ we may want to conclude that we have an *i.i.d* source with a high probability for 1.

The particular application we are motivated by arises in high speed chip-to-chip communications, and is commonly called the backplane channel [9]. Here, residual reflections between inter-chip connects form

a significant source of interference. Because of parasitic capacitances, the channel is highly non-linear as well, and consequently the residual signal that determines the channel state is not a linear function of past inputs as in typical interference channels. We therefore consider a channel model where the output is not necessarily a linear function of the input, and in addition, the channel encountered by any input symbol is determined by the prior outputs. Such a model also yields to analytical transparency. Therefore, we begin with estimation problems in channels whose state is determined by the output memory. See also [10], [11] for other examples of output memory channels.

Our main results are summarized in Section III. These results show how to look at a data sample and identify states of the channel that are amenable to accurate estimation from the sample. They also allow us to sometimes (depending on how the data looks) conclude that certain naive estimators of stationary probabilities or channel transition probabilities happen to be accurate, *even if the channel evolution is slow mixing*. To obtain these results, we combine universal compression results of the context tree weighting algorithm with coupling arguments by Aldous [12].

A. Prior Work on Estimation of Markov Processes

Estimation for Markov processes has been extensively studied and falls into three major categories (i) consistency of estimators *e.g.,* [13], [14], [15], [16], (ii) guarantees on estimates that hold eventually almost surely *e.g.,* [17], [18], and (iii) guarantees that hold for all sample sizes but which depend on the model parameters *e.g.,* [19], [20], [21], [22]. As mentioned earlier, performance of any estimator cannot not be bounded uniformly over all Markov models, something reflected in the line (iii) of research and in our work. The crucial distinction is that we can gauge from the observed sample *if* our estimator is doing well, something that does not hold in (iii). The list is not exhaustive, rather work closest to the approaches we take.

In [19], [20], [22], exponential upper bounds on probability of incorrect estimation of (i) conditional and stationary probabilities and (ii) the underlying context tree, are provided for variants of Rissanen's algorithm *context* and penalized maximum likelihood estimator. However, the introduced deviation bounds depend on the parameters (*e.g.,* depth of the tree, stationary probabilities) of underlying process which are unknown a-priori in practice. In [21], the problem of estimating a stationary ergodic processes by finite memory Markov processes based on an n-length sample of the process is addressed. A measure of distance between the true process and its estimation is introduced and a convergence rate with respect to that measure is provided. However, the deviation bound holds only when the infimum of conditional probabilities of symbols given the pasts are bounded away from zero.

In this work, given a realization of a Markov process, we consider a coarser model and provide deviation bounds for sequences which have occurred frequently enough in the sample. In contrast to prior literature, while we make an assumption justified by the physical model—that dependencies die down in the Markov model class we consider—our bounds can be calculated using only parameters which are well-approximated from data. In particular, we do not assume neither a-prior knowledge on the depth of context tree of the process nor the conditional probabilities given the pasts are bounded uniformly away from zero.

B. Prior Work on Information rates of channel with memory

The computation of the capacity of channels with memory has long been an open problem. The past efforts mainly focus on the class of finite state channels. To summarize, researchers have considered (i) computing information rate, (ii) finding lower and upper bounds for the information rates, and (iii) capacity achieving distributions. A comprehensive review is available in [23], [24], [25], [26]. In particular, for ISI channels with an average power constraint and Gilbert-Elliot-type channels, the capacity is already known. In addition, the capacity for output memory channels with an additive noise channel (independent of input)

was computed in [10]. Note that, in this line of work, the channel model and its properties was assumed to be known.

II. Markov Processes and Channels

A. Alphabet and strings

Most notation here is standard, we include them for completeness. $\mathcal{A}$ is a finite alphabet with cardinality $|\mathcal{A}|$, $\mathcal{A}^* = \bigcup_{k\geq 0} \mathcal{A}^k$ and $\mathcal{A}^\infty$ denotes the set of all semi-infinite strings of symbols in $\mathcal{A}$.

We denote the length of a string $\mathbf{u} = u_1, \ldots, u_l \in \mathcal{A}^l$ by $|\mathbf{u}|$, and use $\mathbf{u}_i^j = (u_i, \cdots, u_j)$. The concatenation of strings $\mathbf{w}$ and $\mathbf{v}$ is denoted by $\mathbf{wv}$. A string $\mathbf{v}$ is a *suffix* of $\mathbf{u}$, denoted by $\mathbf{v} \preceq \mathbf{u}$, if there exists a string $\mathbf{w}$ such that $\mathbf{u} = \mathbf{wv}$. A set $\mathcal{T}$ of strings is *suffix-free* if no string of $\mathcal{T}$ is a suffix of any other string in $\mathcal{T}$.

B. Trees

As in [27] for example, we use full $\mathcal{A}-$ary trees to represent the states of a Markov process. We denote complete trees $\mathcal{T}$ as a suffix-free set $\mathcal{T} \subset \mathcal{A}^*$ of strings (the *leaves*) whose lengths satisfy Kraft's lemma with equality. The depth of the tree $\mathcal{T}$ is defined as $d(\mathcal{T}) = max\{\,|\mathbf{u}| \,:\, \mathbf{u} \in \mathcal{T}\,\}$. A string $\mathbf{v} \in \mathcal{A}^*$ is an *internal node* of $\mathcal{T}$ if either $\mathbf{v} \in \mathcal{T}$ or there exists $\mathbf{u} \in \mathcal{T}$ such that $\mathbf{v} \preceq \mathbf{u}$. The *children* of an internal node $\mathbf{v}$ in $\mathcal{T}$, are those strings (if any) $a\mathbf{v}, a \in \mathcal{A}$ which are themselves either internal nodes or leaves in $\mathcal{T}$.

For any internal node $\mathbf{w}$ of a tree $\mathcal{T}$, let $\mathcal{T}_\mathbf{w} = \{\mathbf{u} \in \mathcal{T} : \mathbf{w} \preceq \mathbf{u}\}$ be the subtree rooted at $\mathbf{w}$. Given two trees $\mathcal{T}_1$ and $\mathcal{T}_2$, we say that $\mathcal{T}_1$ is included in $\mathcal{T}_2$ ($\mathcal{T}_1 \preceq \mathcal{T}_2$), if all the leaves in $\mathcal{T}_1$ are either leaves or internal nodes of $\mathcal{T}_2$.

C. Models

Let $\mathcal{P}^+(\mathcal{A})$ be the set of all probability distributions on $\mathcal{A}$ such that every probability is strictly positive.

Definition 1: A context tree *model* is a finite full tree $\mathcal{T} \subset \mathcal{A}^*$ with a collection of probability distributions $q_\mathbf{s} \in \mathcal{P}^+(\mathcal{A})$ assigned to each $\mathbf{s} \in \mathcal{T}$. We will refer to the elements of $\mathcal{T}$ as *states* (or *contexts*), and $q(\mathcal{T}) = \{q(a|\mathbf{s}) : \mathbf{s} \in \mathcal{T},\ a \in \mathcal{A}\}$ as the set of *state transition probabilities* or the *process parameters*. □

Every model $(\mathcal{T}, q(\mathcal{T}))$ allows for an irreducible, aperiodic[1] and ergodic [28]. Such Markov process has a unique stationary distribution μ satisfying

$$\mu Q = \mu, \tag{1}$$

where Q is the standard transition probability matrix formed using $q(\mathcal{T})$. Let $p_{\mathcal{T},q}$ be the unique stationary Markov process $\{\ldots, Y_0, Y_1, Y_2, \ldots\}$ which takes values in $\mathcal{A}$ satisfying

$$p_{\mathcal{T},q}(Y_1|Y_{-\infty}^0) = q(Y_1|\mathbf{s})$$

whenever $\mathbf{s} = \mathbf{c}_\mathcal{T}(Y_{-\infty}^0)$, where $\mathbf{c}_\mathcal{T} : \mathcal{A}^\infty \to \mathcal{T}$ is the unique suffix $\mathbf{s} \preceq Y_{-\infty}^0$ in $\mathcal{T}$. As a note, when we write out actual strings in transition probabilities as in $q(0|1000)$, the state 1000 is the sequence of bits as we encounter them when reading the string left to right. If 0 follows the state 1100, the next state is a suffix of 11000, and if 1 follows 1100, the next state is a suffix of 11001.

Observation 1: A useful observation is that any model $(\mathcal{T}, q(\mathcal{T}))$ yields the same Markov process as a model $(\mathcal{T}', q(\mathcal{T}'))$ where $\mathcal{T} \preceq \mathcal{T}'$ and for all $\mathbf{s}' \in \mathcal{T}'$, $q(\cdot|\mathbf{s}') = q(\cdot|\mathbf{c}_\mathcal{T}(\mathbf{s}'))$. □

[1] Irreducible since $q_s \in \mathcal{P}^+(\mathcal{A})$, aperiodic since any state $\mathbf{s} \in \mathcal{T}$ can be reached in either $|\mathbf{s}|$ or $|\mathbf{s}|+1$ steps.

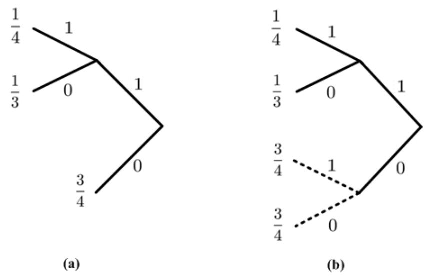

Fig. 1. (a) States and parameters of a Markov process in Example 1, (b) Same Markov process reparameterized to be a complete tree of depth 2. We can similarly reparameterize the process on the left with a complete tree of any depth larger than 2.

Example 1: Let $(\mathcal{T}, q(\mathcal{T}))$ be a Model with $\mathcal{T} = \{11, 01, 0\}$ and $q(1|11) = \frac{1}{4}, q(1|01) = \frac{1}{3}, q(1|0) = \frac{3}{4}$. Fig. 1. (b) shows the Markov process as a model $(\mathcal{T}', q(\mathcal{T}'))$ with $\mathcal{T}' = \{11, 01, 10, 00\}$ satisfying conditions in Property 1. □

D. Channel Model

We focus on Markov channels defined as follows. Both input $\{X_n\}_{n=1}^{\infty}$ and output $\{Y_n\}_{n=1}^{\infty}$ are finite alphabet processes taking values in $\mathcal{A}$ and the state of channel in each instant depend on sequence of prior outputs of the channel. The input process is drawn from an *i.i.d* process, namely $P(X_n = a) = p_a$ for all $n \in \mathbb{N}$ and $a \in \mathcal{A}$, provided that $\sum_{a \in \mathcal{A}} p_a = 1$. We assume that there is no feedback in this channel setup. The joint probability distribution of the channel is

$$P(x_1^n, y_1^n) = \prod_{k=1}^{n} P(x_k) P(y_k | y_{-\infty}^{k-1}, x_k).$$

We consider the process $\{(X_n, Y_n)\}_{n=1}^{\infty}$, and model it as a Markov process $p_{\mathcal{T},q}$. Here, $\mathcal{T}$ will be a complete $|\mathcal{A}|$-ray tree. To every state $\mathbf{s} \in \mathcal{T}$ and $(a, b) \in \mathcal{A} \times \mathcal{A}$, let

$$q_{\mathbf{s}}^{(a,b)} = P(X_k = a, Y_k = b | \mathbf{c}_{\mathcal{T}}(Y_{-\infty}^{k-1}) = \mathbf{s}), \quad \forall k \in \mathbb{N}$$

where $\mathbf{c}_{\mathcal{T}}(y_{-\infty}^{k-1}) \in \mathcal{T}$ (defined above) is the state of the channel at time k. For convenience, we define

$$\theta_{\mathbf{s}}(b|a) = P(Y_1 = b | \mathbf{c}_{\mathcal{T}}(Y_{-\infty}^{0}) = \mathbf{s}, X_1 = a).$$

The set $\Theta_{\mathbf{s}} = \{\theta_{\mathbf{s}}(\cdot|a) \colon a \in \mathcal{A}\}$ is called the set of all conditional probabilities associated with state $\mathbf{s}$. As a remark, for all $a \in \mathcal{A}$ and $\mathbf{s} \in \mathcal{T}$, $\theta_{\mathbf{s}}(\cdot|a) \in \mathcal{P}^{+}(\mathcal{A})$ and corresponds to transition probabilities when the state of channel is $\mathbf{s}$ and input symbol is a. The set $\boldsymbol{\Theta}_{\mathcal{T}} = \{\Theta_{\mathbf{s}} \colon \mathbf{s} \in \mathcal{T}\}$ is the set of all transition probabilities of channel model and is called *channel parameters*. Therefore, we have

$$q_{\mathbf{s}}^{(a,b)} = p_a \, \theta_{\mathbf{s}}(b|a).$$

Since the input is a known *i.i.d* process, estimating $q(\mathcal{T})$ and $\boldsymbol{\Theta}_{\mathcal{T}}$ are completely equivalent and can be obtained from each other using the equations above.

As emphasized in the introduction, we do not assume the true channel model is known nor do we assume it is fast mixing . We would like to know if we can estimate the channel parameters and the stationary probabilities of various states of the channel even when we are in the domain where the mixing has not happened.

III. MAIN RESULTS AND REMARKS

Motivated by the operation and estimation in backplane channels, we consider the following abstraction. We have n *i.i.d* input symbols and their corresponding outputs, and the input/output pairs evolve as described in Section II-D above. There is no feedback and the input is a known *i.i.d* process. The channel input/output pairs then form a Markov process $p_{\mathcal{T},q}$ where both the set of contexts $\mathcal{T}$, and the transition probabilities $q(\mathcal{T})$ (or equivalently the channel parameters $\mathbf{\Theta}_{\mathcal{T}}$, see Section II-D) are unknown.

Using this sample, we want (i) to approximate as best as possible, the parameter set $\mathbf{\Theta}_{\mathcal{T}}$ (ii) the stationary probabilities $\mu(\mathbf{s})$ of observing an output string $\mathbf{s} \in \mathcal{T}$, and (iii) estimate or at least obtain heuristics of the information rate of the process.

Two distinct problems complicate estimation of $\mathbf{\Theta}_{\mathcal{T}}$ and the stationary probabilities. First is the issue that the memory may be too long to handle—in fact, if the source has long enough memory it may not be possible, with n samples, to distinguish the source even from a memoryless source (Example 2). Secondly, even if the source has only one bit of memory, it may be arbitrarily slow mixing as in the example in the introduction and in Example 3. No matter what n is, there will be sources against which our estimates perform very poorly.

A natural way to deal with $\mathcal{T}$ being unknown is to try estimating a potentially coarser approximation $p_{\tilde{\mathcal{T}},\tilde{q}}$ to the true model, where $\tilde{\mathcal{T}} = \mathcal{A}^{k_n}$ for some known k_n. With the benefit of hindsight, we take $k_n = \mathcal{O}(\log n)$ and write $k_n = \alpha_n \log n$ for some function $\alpha_n = \mathcal{O}(1)$. This scaling of k_n also reflects the well known conditions for consistency of estimation of Markov processes in [15].

The input/output pairs obtained by the coarser channel model will be a stationary Markov process. We call it the *aggregation* of $p_{\mathcal{T},q}$ and is specified in Definition 2. The aggregation matches all joint distributions of $p_{\mathcal{T},q}$ involving variables that are less than k_n apart from each other—namely if $i_1, \dots, i_l$ are any increasing sequence of l numbers such that $i_l - i_1 + 1 < k_n$, then $p_{\mathcal{T},q}(X_{i_1}Y_{i_1}, \dots, X_{i_l}, Y_{i_l}) = p_{\tilde{\mathcal{T}},\tilde{q}}(X_{i_1}, Y_{i_1}, \dots, X_{i_l}, Y_{i_l})$. We show that the information rate corresponding to the aggregation $p_{\tilde{\mathcal{T}},\tilde{q}}$ is a lower bound on the information rate of the true channel input/output process $p_{\mathcal{T},q}$ in Theorem 3. While it may not be possible to directly use this Theorem 3, we develop of the notion of a *partial information rate*, a useful heuristic when the source has not mixed. Moreover, Theorem 3 also motivates the estimation questions that form our main results.

The issue remains that the sample of channel input/outputs (given past $Y_{-\infty}^0$) at hand is from $p_{\mathcal{T},q}$, *not* from $p_{\tilde{\mathcal{T}},\tilde{q}}$. To obtain the parameters of the aggregated model, $\tilde{\mathbf{\Theta}}_{\tilde{\mathcal{T}}}$, we use a *naive* estimator—we simply pretend that our sample was in fact from $p_{\tilde{\mathcal{T}},\tilde{q}}$. Equivalently, we assume that for any output sequence $\mathbf{w} \in \tilde{\mathcal{T}}$, the subsequence of output symbols in the sample that follow $\mathbf{w}$ and associated with the same input letter a is *i.i.d*. For example, if our sample has output 1101010100 for input 1111100000, and $\mathbf{w} = 10$, the output subsequence following $\mathbf{w}$ is 1110, and these outputs correspond to inputs 1000 respectively. We would then simply estimate $\tilde{\theta}_{10}(1|1) = 1$ and $\tilde{\theta}_{10}(1|0) = 2/3$. If $p_{\mathcal{T},q}$ were arbitrary, there is no reason that this naive approach has to be accurate.

However, it is reasonable given our physical motivation that the influence of prior outputs dies down as we look further into the past. We formalize this notion in (3) with a function $f(i)$ that controls how symbols i locations apart can influence each other, and require $\sum_{i=1}^{\infty} f(i) < \infty$. With this assumption, we obtain $\tilde{G} \subseteq \tilde{\mathcal{T}}$, a set of *good states* or *good* strings of channel outputs (strings in $\tilde{\mathcal{T}}$ may not be states of $p_{\mathcal{T},q}$, but we abuse notation here for convenience) from the sample, and show that with probability $\geq 1 - \frac{1}{2^{|\mathcal{A}|^{k_n+1} \log n}}$ (conditioned on any past $Y_{-\infty}^0$), for all $a \in \mathcal{A}$ and $\mathbf{s} \in \tilde{G}$ simultaneously

$$||\tilde{\theta}_{\mathbf{s}}(\cdot|a) - \hat{\theta}_{\mathbf{s}}(\cdot|a)||_1 \leq 2\sqrt{\frac{\ln 2}{\log n} + \frac{\ln 2}{\log \frac{1}{\nu_{k_n}}}}$$

where $\nu_{k_n} = \sum_{i \geq k_n} f(i)$ and $\hat{\theta}_{\mathbf{s}}(\cdot|a)$ is the naive estimator of $\tilde{\theta}_s(\cdot|a)$ described above. In Theorem 7 we

strengthen the convergence rate to polynomial in n if the dependencies die down exponentially.

The above estimation result associated with $\mathbf{s} \in \tilde{G}$ relies not on mixing but on the fact that Markov sources with memory k_n can be universally compressed. Therefore, it is still definitely possible that the counts of $\mathbf{s} \in \tilde{G}$ are nowhere near their stationary probabilities. How do we then interpret their counts? To answer this question, we define a parameter $\eta_{\tilde{G}}$ in (7) that is calculated using just the parameters associated with $\mathbf{s} \in \tilde{G}$. Suppose $\{\nu_i\}$ is summable as well, and let $\Phi_j = \sum_{i \geq j} \nu_i$. A coupling argument on a natural Doob Martingale construction then proves that if the evolution of $p_{\mathcal{T},q}$ restricted to just states in $\tilde{G}$ is aperiodic, then for all $t > 0$, $Y_{-\infty}^0$ and $\mathbf{w} \in \tilde{G}$ the counts of $\mathbf{w}$ in the sample, $N_n(\mathbf{w})$, concentrates:

$$p_{\mathcal{T},q}(|N_n(\mathbf{w}) - \tilde{n}\frac{\mu(\mathbf{w})}{\mu(\tilde{G})}| \geq t|Y_{-\infty}^0) \leq 2\exp\left(-\frac{t^2}{2\tilde{n}}\left(\frac{\eta_{\tilde{G}}^{k_n}(1-\Phi_{k_n})}{4\ell_n + \eta_{\tilde{G}}^{k_n}(1-\Phi_{k_n})}\right)^2\right)$$

where ℓ_n is the smallest integer such that $\Phi_{\ell_n} \leq \frac{1}{n}$, $\tilde{n}$ is the total count of good states in the sample and μ denotes the stationary distribution of $p_{\mathcal{T},q}$.

IV. Background

A. Context Tree Weighting

Context tree weighting algorithm is a universal data compression algorithm for Markov sources [27], [29], and the algorithm can be used to capture several insights into how Markov processes behave in the non-asymptotic regime. Let y_1^n be sequence of symbols from $\mathcal{A}$. Let $\hat{\mathcal{T}} = \mathcal{A}^D$ for some positive integer D. For all $\mathbf{s} \in \hat{\mathcal{T}}$ and $a \in \mathcal{A}$, let $n_{\mathbf{s}}^a$ be the empirical counts of string $\mathbf{s}a$ in y_1^n. The depth-D context tree weighting constructs a distribution

$$p_c(y_1^n|y_{-D}^0) \geq 2^{-|\mathcal{A}|^{D+1}\log n} \prod_{\mathbf{s}\in\hat{\mathcal{T}}} \prod_{a\in\mathcal{A}} \left(\frac{n_{\mathbf{s}}^a}{\sum_{a\in\mathcal{A}} n_{\mathbf{s}}^a}\right)^{n_{\mathbf{s}}^a}.$$

Note that no Markov source with memory D could have given a higher probability to y_1^n than

$$\prod_{\mathbf{s}\in\hat{\mathcal{T}}} \prod_{a\in\mathcal{A}} \left(\frac{n_{\mathbf{s}}^a}{\sum_{a\in\mathcal{A}} n_{\mathbf{s}}^a}\right)^{n_{\mathbf{s}}^a}.$$

So, if $|A|^D \log n = o(n)$, then p_c underestimates any memory-D Markov probability by only a subexponential factor. Therefore, $D = \mathcal{O}(\log n)$ is going to be the case of particular interest.

B. Coupling for Markov Processes

Coupling is an elegant technique that will help us understand how the counts of certain strings in a sample behave. Let $p_{\mathcal{T},q}$ be our Markov source. A coupling ω for $p_{\mathcal{T},q}$ is a joint process $\{Y_n, \bar{Y}_n\}$ such which both $\{Y_n\}$ and $\{\bar{Y}_n\}$ are marginally faithful evolutions of $p_{\mathcal{T},q}$. Say $\{Y_n\}$ and $\{\bar{Y}_n\}$ here are copies of $p_{\mathcal{T},q}$ that were started with states $\mathbf{s}, \mathbf{s}' \in \mathcal{T}$, respectively. The joint distribution $\omega(\{Y_n, \bar{Y}_n\})$ is otherwise arbitrary, but we want to encourage them to *coalesce*. For $\mathbf{w} \in \mathcal{T}$, $N_n(\mathbf{w})$ ($\bar{N}_n(\mathbf{w})$) is the number of occurrences of $\mathbf{w}$ in $\{Y_n\}$ ($\{\bar{Y}_n\}$). Then, $|\mathbb{E}_{p_{\mathcal{T},q}} N_n(\mathbf{w}) - \mathbb{E}_{p_{\mathcal{T},q}} \bar{N}_n(\mathbf{w})|$ equals

$$\begin{aligned}\left|\sum_{i=1}^n E_\omega\left[\mathbb{1}\left(\mathbf{c}_{\mathcal{T}}(Y_{-\infty}^i) = \mathbf{w}\right) - \mathbb{1}\left(\mathbf{c}_{\mathcal{T}}(\bar{Y}_{-\infty}^i) = \mathbf{w}\right)\right]\right| &\leq \sum_{i=1}^n \left|\mathbb{E}_\omega\left[\mathbb{1}\left(\mathbf{c}_{\mathcal{T}}(Y_{-\infty}^i) = \mathbf{w}\right) - \mathbb{1}\left(\mathbf{c}_{\mathcal{T}}(\bar{Y}_{-\infty}^i) = \mathbf{w}\right)\right]\right| \\ &\leq \sum_{i=1}^n \omega\left(\mathbf{c}_{\mathcal{T}}(Y_{-\infty}^i) \neq \mathbf{c}_{\mathcal{T}}(\bar{Y}_{-\infty}^i)\right).\end{aligned}$$

For tutorials, see *e.g.,* [30], [8], [31].

V. Long Memory and Slow Mixing

There are two distinct difficulties in estimating Markov processes as the ones we are interested in. The first is memory that is too long to handle given the size of the sample at hand. The second issue is that even though the underlying process might be ergodic, the transition probabilities are so small such that the process effectively acts like a non-ergodic process given the sample size available. We illustrate these problems in following simple examples.

Example 2: Let $\mathcal{T} = \mathcal{A}^k$ denote a full tree with depth k and $\mathcal{A} = \{0,1\}$. Assume that $q(1|0^k) = 2\epsilon$ and $q(1|10^{k-1}) = 1-\epsilon$ with $\epsilon > 0$, and let $q(1|\mathbf{s}) = \frac{1}{2}$ (where 0^k indicates a string with k consecutive zeros) for all other $\mathbf{s} \in \mathcal{T}$. Let $p_{\mathcal{T},q}$ represent the stationary ergodic Markov process associated with this model. Observe that stationary probability of being in state 0^k is $\frac{1}{2^{k+1}-1}$ while all other states have stationary probability $\frac{2}{2^{k+1}-1}$. Let Y_1^n be a realization of this process with initial state $1^k \preceq Y_{-\infty}^0$. Suppose $k \gg \omega(\log n)$. [2] With high probability we will never find a string of $k-1$ zeros among n samples, and every bit is generated with probability 1/2. Thus with this sample size, with high probability, we cannot distinguish $p_{\mathcal{T},q}$ from an *i.i.d* Bernoulli(1/2) process. □

We therefore require that dependencies die down by requiring that channel parameters $\theta^a_{1\mathbf{s}}$ and $\theta^a_{0\mathbf{s}}$, corresponding to sibling contexts $1\mathbf{s}$ and $0\mathbf{s}$, satisfy (3) in Section VI.

Example 3: Let $\mathcal{A} = \{0,1\}$ and $\mathcal{T} = \{0,1\}$ with $q(1|1) = 1-\epsilon$, and $q(1|0) = \epsilon$. For $\epsilon > 0$, this model represents a stationary ergodic Markov processes with stationary distributions $\mu(1) = \frac{1}{2}, \mu(0) = \frac{1}{2}$. Let $\mathcal{T}' = \{0,1\}$ with $q'(1|1) = 1-\epsilon, q'(1|0) = 2\epsilon$. Similarly, for $\epsilon > 0$ this model represents a stationary ergodic Markov processes with stationary distributions $\mu'(1) = \frac{2}{3}, \mu'(0) = \frac{1}{3}$. Suppose we have a length-n sample. In this case, we cannot distinguish between these two models if $\epsilon \ll o(1/n)$. □

A. Lower Bound on Information Rate

Consider a channel with state tree $\mathcal{T}$ and parameter set $\mathbf{\Theta}_{\mathcal{T}}$. Suppose that $d(\mathcal{T}) = D < \infty$. The information rate for an *i.i.d* input process with $P(X_k = a) = p_a$ for this channel is

$$
\begin{aligned}
R_{\mathcal{T}} &\stackrel{\text{def}}{=} \lim_{n\to\infty} \frac{1}{n} I(X^n;Y^n) \\
&= \lim_{n\to\infty} \frac{1}{n}\big[H(Y^n) - H(Y^n|X^n)\big] \\
&\stackrel{a}{=} \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}\Big[H(Y_k|Y^{k-1}) - H(Y_k|Y^{k-1},X_k)\Big] \\
&\stackrel{b}{=} \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{D}\sum_{y^{k-1}} P(Y^{k-1}=y^{k-1})\Big[H(Y_k|Y^{k-1}=y^{k-1}) - H(Y_k|Y^{k-1}=y^{k-1},X_k)\Big] \\
&\qquad + \lim_{n\to\infty} \frac{1}{n}\sum_{k=D+1}^{n}\sum_{y^{k-1}} P(Y^{k-1}=y^{k-1})\Big[H(Y_k|Y^{k-1}=y^{k-1}) - H(Y_k|Y^{k-1}=y^{k-1},X_k)\Big] \\
&\stackrel{c}{=} \lim_{n\to\infty} \frac{1}{n}\sum_{k=D+1}^{n}\sum_{\mathbf{s}\in\mathcal{T}}\sum_{\substack{y^{k-1}:\\ \mathbf{s}\preceq y^{k-1}}} P(Y^{k-1}=y^{k-1})\Big[H(Y_k|Y^{k-1}=y^{k-1}) - H(Y_k|Y^{k-1}=y^{k-1},X_k)\Big]
\end{aligned}
$$

where (a) is by chain rule for entropy and the fact that input process is independent of output process. This condition always is correct as long as there is no feedback in the channel and (b) is straightforward

[2] A function $f_n = \omega(g_n)$ if $\lim_{n\to\infty} f_n/g_n = \infty$.

from definition of the conditional entropy. The equality in (c) holds since the first term in (b) vanishes as $n \to \infty$ and observing that for $k \geq D+1$,

$$\mathcal{A}^{k-1} = \bigcup_{\mathbf{s}\in\mathcal{T}} \{y^{k-1} \in \mathcal{A}^{k-1} \colon \mathbf{s} \preceq y^{k-1}\}.$$

Note that for all $k \in \mathbb{N}$, if $\mathbf{s} \preceq y^{k-1}$, it can easily be shown that

$$\begin{aligned} H(Y_k|Y^{k-1} = y^{k-1}) - H(Y_k|Y^{k-1} = y^{k-1}, X_k) &= \sum_{a\in\mathcal{A}} p_a \sum_{b\in\mathcal{A}} \theta_{\mathbf{s}}(b|a) \log \frac{\theta_{\mathbf{s}}(b|a)}{\sum_{a'\in\mathcal{A}} p_{a'}\theta_{\mathbf{s}}(b|a')} \\ &\stackrel{\text{def}}{=} R_{\mathbf{s}}(\Theta_{\mathbf{s}}). \end{aligned}$$

Therefore,

$$\begin{aligned} R_{\mathcal{T}} &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=D+1}^{n} \sum_{\mathbf{s}\in\mathcal{T}} \sum_{\substack{y^{k-1}: \\ \mathbf{s}\preceq y^{k-1}}} P(Y^{k-1} = y^{k-1}) R_{\mathbf{s}}(\Theta_{\mathbf{s}}) \\ &\stackrel{d}{=} \sum_{\mathbf{s}\in\mathcal{T}} R_{\mathbf{s}}(\Theta_{\mathbf{s}}) \left[\lim_{n\to\infty} \frac{1}{n} \sum_{k=D+1}^{n} \sum_{\substack{y^{k-1}: \\ \mathbf{s}\preceq y^{k-1}}} P(Y^{k-1} = y^{k-1}) \right] \\ &\stackrel{e}{=} \sum_{\mathbf{s}\in\mathcal{T}} R_{\mathbf{s}}(\Theta_{\mathbf{s}}) \left[\lim_{k\to\infty} \sum_{\substack{y^{k-1}: \\ \mathbf{s}\preceq y^{k-1}}} P(Y^{k-1} = y^{k-1}) \right] \\ &\stackrel{f}{=} \sum_{\mathbf{s}\in\mathcal{T}} \mu(\mathbf{s}) R_{\mathbf{s}}(\Theta_{\mathbf{s}}). \end{aligned}$$

The equality in (d) is by changing the order of summations and (e) follows by using Cesàro's lemma [32]. Finally, (f) holds by properties of stationary distribution in Markov processes. As a remark, note that for fixed input distribution, $R_{\mathbf{s}}$ is a function of $\Theta_{\mathbf{s}} = \{\theta_{\mathbf{s}}(\cdot|a) \colon a \in \mathcal{A}\}$.

Lemma 1: For fixed input distribution, $R_{\mathbf{s}}(\Theta_{\mathbf{s}})$ is convex in $\Theta_{\mathbf{s}}$.

Proof Let $\lambda \in [0,1]$ and $\overline{\lambda} = 1-\lambda$. Let $\Theta_{\mathbf{s}} = \{\theta_{\mathbf{s}}(\cdot|a) \colon a \in \mathcal{A}\}$ and $\Theta'_{\mathbf{s}} = \{\theta'_{\mathbf{s}}(\cdot|a) \colon a \in \mathcal{A}\}$ be two sets of valid conditional distributions associated to state $\mathbf{s}$. Then,

$$\begin{aligned} &R_{\mathbf{s}}(\lambda\Theta_{\mathbf{s}} + \overline{\lambda}\Theta'_{\mathbf{s}}) = \\ &\quad = \sum_{a\in A} p_a \sum_{b\in\mathcal{A}} \left[\Big(\lambda\theta_{\mathbf{s}}(b|a) + \overline{\lambda}\theta'_{\mathbf{s}}(b|a)\Big) \log \frac{\lambda\theta_{\mathbf{s}}(b|a) + \overline{\lambda}\theta'_{\mathbf{s}}(b|a)}{\sum\limits_{a'\in\mathcal{A}} p_{a'} \Big(\lambda\theta_{\mathbf{s}}(b|a') + \overline{\lambda}\theta'_{\mathbf{s}}(b|a')\Big)} \right] \\ &\quad = \sum_{a\in A} p_a \sum_{b\in\mathcal{A}} \left[\Big(\lambda\theta_{\mathbf{s}}(b|a) + \overline{\lambda}\theta'_{\mathbf{s}}(b|a)\Big) \log \frac{\lambda\theta_{\mathbf{s}}(b|a) + \overline{\lambda}\theta'_{\mathbf{s}}(b|a)}{\lambda \sum\limits_{a'\in\mathcal{A}} p_{a'}\theta_{\mathbf{s}}(b|a') + \overline{\lambda} \sum\limits_{a'\in\mathcal{A}} p_{a'}\theta'_{\mathbf{s}}(b|a')} \right] \\ &\quad \stackrel{e}{\leq} \sum_{a\in A} p_a \sum_{b\in\mathcal{A}} \left[\lambda\,\theta_{\mathbf{s}}(b|a) \log \frac{\lambda\cdot\theta_{\mathbf{s}}(b|a)}{\lambda\cdot \sum\limits_{a'\in\mathcal{A}} p_{a'}\theta_{\mathbf{s}}(b|a')} + \overline{\lambda}\,\theta'_{\mathbf{s}}(b|a) \log \frac{\overline{\lambda}\cdot\theta'_{\mathbf{s}}(b|a)}{\overline{\lambda}\cdot \sum\limits_{a'\in\mathcal{A}} p_{a'}\theta'_{\mathbf{s}}(b|a')} \right] \\ &\quad = \lambda R_{\mathbf{s}}(\Theta_{\mathbf{s}}) + \overline{\lambda} R_{\mathbf{s}}(\Theta'_{\mathbf{s}}) \end{aligned}$$

where the inequality in (e) follows by using *log-sum* inequality (see *e.g.,* [32]). $\square$

Since the memory is unknown a-priori, a natural approach, known to be consistent, is to use a potentially coarser model with depth k_n. Here, k_n increases logarithmically with the sample size n, and reflects [15]

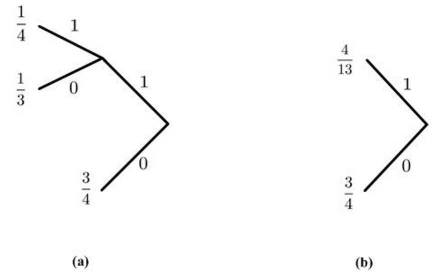

Fig. 2. (a) Markov process in Example 4, (b) Aggregated model at depth 1. From Observation 1, the model on the left can be reparameterized to be a complete tree at any depth ≥ 2. We can hence ask for its aggregation at any depth. Aggregations of the above model on the left at depths ≥ 2 will hence be the model itself.

well known results on consistent estimation of Markov processes. We show that coarser models formed by properly aggregating states of the original channel are useful in lower bounding information rates of the true channel.

Definition 2: We say that $p_{\tilde{\mathcal{T}},\tilde{q}}$ *aggregates* $p_{\mathcal{T},q}$ (or $p_{\mathcal{T},q}$ *refines* $p_{\tilde{\mathcal{T}},\tilde{q}}$), if $\tilde{\mathcal{T}} \preceq \mathcal{T}$ and $p_{\tilde{\mathcal{T}},\tilde{q}}$ be the stationary Markov process with state transition probabilities given by

$$\tilde{q}(a|\mathbf{w}) = \frac{\sum_{\mathbf{v}\in\mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v})q(a|\mathbf{v})}{\sum_{\mathbf{v}'\in\mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v}')}$$

for all $\mathbf{w} \in \tilde{\mathcal{T}}$ and $a \in \mathcal{A}$, where μ is the stationary distribution associated with $p_{\mathcal{T},q}$. Using Observation 1, wolog, no matter what $\tilde{\mathcal{T}}$ is, we will assume $p_{\mathcal{T},q}$ has states $\mathcal{T}$ such that $\tilde{\mathcal{T}} \preceq \mathcal{T}$. $\square$

Example 4: Let $p_{\mathcal{T},q}$ be a Markov process with $\mathcal{T} = \{11, 01, 0\}$ and $q(1|11) = \frac{1}{4}, q(1|01) = \frac{1}{3}, q(1|0) = \frac{3}{4}$. For this model, we have $\mu(11) = \frac{4}{25}, \mu(01) = \frac{9}{25}$ and $\mu(0) = \frac{12}{25}$. Fig. 2. (b) shows an aggregated process $p_{\tilde{\mathcal{T}},\tilde{q}}$ with $\tilde{\mathcal{T}} = \{1, 0\}$. Notice that $\tilde{q}(1|1) = \left(\frac{4}{25}\frac{1}{4} + \frac{9}{25}\frac{1}{3}\right)/\left(\frac{4}{25} + \frac{9}{25}\right) = \frac{4}{13}$ and $\tilde{q}(1|0) = \frac{3}{4}$. $\square$

Lemma 2: Let $p_{\mathcal{T},q}$ be a stationary Markov process with stationary distribution μ. If $p_{\tilde{\mathcal{T}},\tilde{q}}$ aggregates $p_{\mathcal{T},q}$ then it has a unique stationary distribution $\tilde{\mu}$ and for every $\mathbf{w} \in \tilde{\mathcal{T}}$

$$\tilde{\mu}(\mathbf{w}) = \sum_{\mathbf{v}\in\mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v}).$$

Moreover, for all $a_1^k \in \mathcal{A}^k$ such that a_1^k is an internal node of $\tilde{\mathcal{T}}$ we have $\tilde{\mu}(a_1^k) = \mu(a_1^k)$.

Proof Let $\tilde{Q}$ be the transition probability matrix formed by the states of $p_{\tilde{\mathcal{T}},\tilde{q}}$. First notice that by definition, for all $\mathbf{w}, \mathbf{w}' \in \tilde{\mathcal{T}}$

$$\tilde{Q}(\mathbf{w}|\mathbf{w}') = \begin{cases} 0 & \text{if } \nexists a \in \mathcal{A} \quad \text{s.t. } \mathbf{w} \preceq \mathbf{w}'a \\ \tilde{q}(a|\mathbf{w}') & \text{if } \mathbf{w} \preceq \mathbf{w}'a \quad \text{for some } a \in \mathcal{A} \end{cases}$$

Since $p_{\mathcal{T},q}$ is irreducible and aperiodic, $p_{\tilde{\mathcal{T}},\tilde{q}}$ will also be irreducible and aperiodic and thus, has a unique stationary distribution $\tilde{\mu}$. Hence, there exists a unique solution for

$$\tilde{\mu}(\mathbf{w}) = \sum_{\mathbf{w}'\in\tilde{\mathcal{T}}} \tilde{\mu}(\mathbf{w}')\tilde{Q}(\mathbf{w}|\mathbf{w}') \quad \forall \mathbf{w} \in \tilde{\mathcal{T}}. \tag{2}$$

We will consider a candidate solution of the form

$$\tilde{\mu}(\mathbf{w}) = \sum_{\mathbf{v}\in\mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v})$$

for every $\mathbf{w} \in \tilde{\mathcal{T}}$ and show that this candidate will satisfy (2). Then, the claim will follow by uniqueness of the solution. To show this, note that for $\forall \mathbf{w} \in \tilde{\mathcal{T}}$

$$\begin{aligned}
\sum_{\substack{\mathbf{w}' \in \tilde{\mathcal{T}} \\ \mathbf{w} \preceq \mathbf{w}' a}} \tilde{\mu}(\mathbf{w}') \tilde{q}(a|\mathbf{w}') &= \sum_{\substack{\mathbf{w}' \in \tilde{\mathcal{T}} \\ \mathbf{w} \preceq \mathbf{w}' a}} \left[\sum_{\mathbf{v} \in \mathcal{T}_{\mathbf{w}'}} \mu(\mathbf{v}) \right] \frac{\sum_{\mathbf{v} \in \mathcal{T}_{\mathbf{w}'}} \mu(\mathbf{v}) q(a|\mathbf{v})}{\sum_{\mathbf{v}' \in \mathcal{T}_{\mathbf{w}'}} \mu(\mathbf{v}')} \\
&= \sum_{\substack{\mathbf{w}' \in \tilde{\mathcal{T}} \\ \mathbf{w} \preceq \mathbf{w}' a}} \sum_{\mathbf{v} \in \mathcal{T}_{\mathbf{w}'}} \mu(\mathbf{v}) q(a|\mathbf{v}) \\
&= \sum_{\substack{\mathbf{w}' \in \tilde{\mathcal{T}} \\ \mathbf{w} \preceq \mathbf{w}' a}} \sum_{\mathbf{v} \in \mathcal{T}_{\mathbf{w}'}} \mu(\mathbf{v} a) \\
&\overset{d}{=} \sum_{\mathbf{s} \in \mathcal{T}_{\mathbf{w}}} \mu(\mathbf{s}) \\
&= \tilde{\mu}(\mathbf{w})
\end{aligned}$$

where (d) follows by observing that

$$\mathbf{w} \preceq \left\{ \mathbf{v} a \colon \exists\, \mathbf{w}' \in \tilde{\mathcal{T}}, a \in \mathcal{A} \quad \text{s.t.} \quad \mathbf{w} \preceq \mathbf{w}' a \ \text{ and } \ \mathbf{v} \in \mathcal{T}_{\mathbf{w}'} \right\},$$

and then using properties of the stationary distribution of $p_{\mathcal{T},q}$. Note that the second statement of Lemma automatically follows from the uniqueness of stationary distributions. $\square$

In a similar manner as Definition 2, given any input output process for a channel we can define an aggregated channel with tree $\tilde{\mathcal{T}}$ and parameter set $\tilde{\boldsymbol{\Theta}}_{\tilde{\mathcal{T}}}$. For all $\mathbf{w} \in \tilde{\mathcal{T}}$, let $\tilde{\Theta}_{\mathbf{w}} = \{\tilde{\theta}_{\mathbf{w}}(\cdot|a) : a \in \mathcal{A}\}$ in which for fixed $a \in \mathcal{A}$, $\tilde{\theta}_{\mathbf{w}}(\cdot|a)$ is given by

$$\tilde{\theta}_{\mathbf{w}}(b|a) = \frac{\sum_{\mathbf{v} \in \mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v}) \theta_{\mathbf{v}}(b|a)}{\sum_{\mathbf{v}' \in \mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v}')}, \quad \forall b \in \mathcal{A}$$

Theorem 3: Consider a channel with tree $\mathcal{T}$ and parameter set $\boldsymbol{\Theta}_{\mathcal{T}}$. If $\tilde{\mathcal{T}}$ aggregates $\mathcal{T}$, then $R_{\tilde{\mathcal{T}}} \leq R_{\mathcal{T}}$.

Proof Note that for all $\mathbf{w} \in \tilde{\mathcal{T}}$, $\mathcal{T}_{\mathbf{w}} = \{\mathbf{s} \in \mathcal{T} : \mathbf{w} \preceq \mathbf{s}\}$. Since $\tilde{\mathcal{T}} \preceq \mathcal{T}$, we have

$$\begin{aligned}
R_{\tilde{\mathcal{T}}} &= \sum_{\mathbf{w} \in \tilde{\mathcal{T}}} \tilde{\mu}(\mathbf{w}) R_{\mathbf{w}}(\tilde{\Theta}_{\mathbf{w}}) \\
&\overset{a}{\leq} \sum_{\mathbf{w} \in \tilde{\mathcal{T}}} \tilde{\mu}(\mathbf{w}) \sum_{\mathbf{v} \in \mathcal{T}_{\mathbf{w}}} \left[\frac{\mu(\mathbf{v})}{\sum_{\mathbf{v}' \in \mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v}')} R_{\mathbf{v}}(\Theta_{\mathbf{v}}) \right] \\
&\overset{b}{=} \sum_{\mathbf{w} \in \tilde{\mathcal{T}}} \sum_{\mathbf{v} \in \mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v}) R_{\mathbf{v}}(\Theta_{\mathbf{v}}) \\
&= \sum_{\mathbf{s} \in \mathcal{T}} \mu(\mathbf{s}) R_{\mathbf{s}}(\Theta_{\mathbf{s}}) \\
&= R_{\mathcal{T}}
\end{aligned}$$

where the inequality in (a) holds by Lemma 1 and the fact that $\forall a, b \in \mathcal{A}$

$$\tilde{\theta}_{\mathbf{w}}(b|a) = \frac{\sum_{\mathbf{v} \in \mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v}) \theta_{\mathbf{v}}(b|a)}{\sum_{\mathbf{v}' \in \mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v}')}.$$

The equality in (b) hods since $\tilde{\mu}(\mathbf{w}) = \sum_{\mathbf{v}' \in \mathcal{T}_{\mathbf{w}}} \mu(\mathbf{v}')$. $\square$

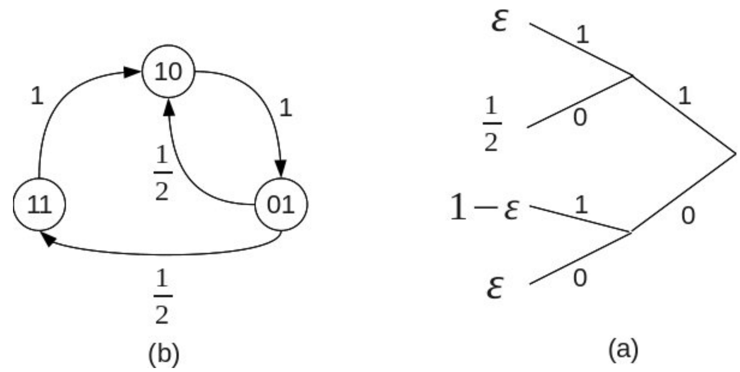

Fig. 3. (a) Markov in Example 5, (b) Same process when $\epsilon = 0$.

Remark In this paper, we are particularly concerned with the slow mixing regime. As our results will show, in general it is not possible to obtain a simple lower bound on the information rate using the data and taking recourse to the Theorem above. Instead, we introduce the *partial* information rate that can be reliably obtained from the data

$$R^p_{\tilde{G}} = \sum_{\mathbf{s}\in\tilde{G}} \frac{\mu(\mathbf{s})}{\mu(\tilde{G})} R_{\mathbf{s}}(\tilde{\Theta}_s),$$

where $\tilde{G} \subseteq \tilde{\mathcal{T}}$ will be a set of *good* states that we show how to identify. The partial information rate is not necessarily a lower bound, but in slow mixing cases it is sometimes the best heuristic possible. We systematically handle the information rates of slow mixing channels using the estimation results below in different paper. □

Notwithstanding the previous remark, we will focus on estimating the aggregated parameters $\mathbf{\Theta}_{\mathcal{T}}$, where $\tilde{\mathcal{T}} = \mathcal{A}^{k_n}$ has depth k_n where k_n grows logarithmically as $\alpha_n \log n$ for some $\alpha_n = \mathcal{O}(1)$. Now $p_{\tilde{\mathcal{T}},\tilde{q}}$ is unknown and we do not have access to samples from it. And there is, of course, no guarantee that the counts of short strings are any more reliable in a long-memory, slow mixing process.

Example 5: Let $\mathcal{T} = \{11, 01, 10, 00\}$ with $q(1|11) = \epsilon$, $q(1|01) = \frac{1}{2}$, $q(1|10) = 1 - \epsilon$, $q(1|00) = \epsilon$. If $\epsilon > 0$, then $p_{\mathcal{T},q}$ is a stationary ergodic binary Markov process. Let μ denote the stationary distribution of this process, respectively. A simple computation shows that $\mu(11) = \frac{1}{7-6\epsilon}$, $\mu(01) = \frac{2-2\epsilon}{7-6\epsilon}$, $\mu(10) = \frac{2-2\epsilon}{7-6\epsilon}$ and $\mu(00) = \frac{2-2\epsilon}{7-6\epsilon}$, and $\mu(1) = \frac{1}{7-6\epsilon} + \frac{2-2\epsilon}{7-6\epsilon} = \frac{3-2\epsilon}{7-6\epsilon}$ and $\mu(0) = \frac{2-2\epsilon}{7-6\epsilon} + \frac{2-2\epsilon}{7-6\epsilon} = \frac{4-4\epsilon}{7-6\epsilon}$.

Suppose we have a length n sample. If $\epsilon \ll \frac{1}{n}$, then $\mu(1) \approx \frac{3}{7}$ and $\mu(0) \approx \frac{4}{7}$ respectively. If the initial state belongs to $\{11, 01, 10\}$, the state 00 will not be visited with high probability in n samples, and it can be seen that the counts of 1 or 0 will not be near the stationary probabilities $\mu(1)$ or $\mu(0)$. For this sample size, the process effectively acts like the irreducible, aperiodic Markov chain in Fig. 3. (b) which can be shown to be fast mixing. Therefore, the stationary probabilities of the chain in Fig. 3. (b), $\frac{\mu(01)}{\mu(1)+\mu(01)}$, $\frac{\mu(10)}{\mu(1)+\mu(01)}$ and $\frac{\mu(11)}{\mu(1)+\mu(01)}$ converge quicker than $\mu(1)$ or $\mu(0)$. □

VI. Estimation of Channel Properties

As noted before in Example 2, if the dependencies could be arbitrary in a channel model, we will not estimate the model accurately no matter how large the sample is. Keeping in mind Observation 1, we formalize dependencies dying down by means of a function $f : \mathbb{Z}^+ \to \mathbb{R}^+$ with $\sum_{i=1}^{\infty} f(i) < \infty$ and

assuming that for all $\mathbf{w} \in \mathcal{A}^*$ and all $c, c' \in \mathcal{A}$

$$|\frac{\theta_{c\mathbf{w}}(b|a)}{\theta_{c'\mathbf{w}}(b|a)} - 1| \leq f(|\mathbf{w}|) \tag{3}$$

where $a, b \in \mathcal{A}$ and $\theta_{c\mathbf{w}}(b|a) = P(Y_1 = b|\mathbf{c}_{\mathcal{T}}(Y_{-\infty}^0) = c\mathbf{w}, X_1 = a)$. Note that the tree is finite iff there exist a finite D such that $f(i) = 0$ for all $i > D$.

As mentioned in the last section, we will focus on set of the aggregated parameters at depth k_n, $\Theta_{\mathcal{A}^{k_n}}$ where $k_n = \alpha_n \log n$. If k_n is large enough, these aggregated parameters start to reflect the underlying parameters $\mathbf{\Theta}_{\mathcal{T}}$. Indeed, by using an elementary argument we will show that both the underlying and aggregated parameters will then be close to the empirically observed values for states that occur frequently enough.

Throughout this Section, we assume that we start with some past $Y_{-\infty}^0$, and we see n samples (X_1^n, Y_1^n) from the channel. All confidence probabilities are conditional probabilities on Y_1^n given X_1^n and $Y_{-\infty}^0$, but we do not write $Y_{-\infty}^0$ out explicitly to avoid cluttering notation. The results hold for all $Y_{-\infty}^0$ (not just with probability 1).

Lemma 4: Let $\{f(i)\}_{i=1}^{\infty}$ be a sequence of real numbers such that there exists some $n_0 \in \mathbb{N}$ for which, $0 \leq f(i) \leq 1$ for all $i \geq n_0$. Then, $\forall j \geq n_0$, we have

$$1 - \sum_{i \geq j} f(i) \leq \prod_{i \geq j}(1 - f(i)) \leq \frac{1}{\prod_{i \geq j}(1 + f(i))}$$

Proof Let E_i be a sequence of independent events such that $\mathbb{P}(E_i) = f(i)$. Then, $\forall j \in \mathbb{N}$ such that $j \geq n_0$, we have

$$1 - \prod_{i \geq j}(1 - f(i)) = \mathbb{P}(\bigcup_{i \geq j} E_i) \leq \sum_{i \geq j} \mathbb{P}(E_i) = \sum_{i \geq j} f(i).$$

Hence,

$$1 - \sum_{i \geq j} f(i) \leq \prod_{i \geq j}(1 - f(i)).$$

Since by assumption $0 \leq f(i) \leq 1$ for all $i \geq n_0$, the second inequality can easily be derived by the fact that

$$\prod_{i \geq j}(1 - f^2(i)) \leq 1.$$

Therefore,

$$\prod_{i \geq j}(1 - f(i)) \leq \frac{1}{\prod_{i \geq j}(1 + f(i))}.$$

□

Lemma 5: Let $\mathcal{T}$ be a model associated with the channel and satisfying condition (3). Suppose $\tilde{\mathcal{T}} \preceq \mathcal{T}$ with $d(\tilde{\mathcal{T}}) = k_n$. If $\sum_{i \geq k_n} f(i) \leq 1$, then for all $\mathbf{w} \in \tilde{\mathcal{T}}$ and $a, b \in \mathcal{A}$

$$\left(1 - \sum_{i \geq k_n} f(i)\right) \max_{\mathbf{s} \in \mathcal{T}_{\mathbf{w}}} \theta_{\mathbf{s}}(b|a) \leq \tilde{\theta}_{\mathbf{w}}(b|a) \leq \frac{\min_{\mathbf{s} \in \mathcal{T}_{\mathbf{w}}} \theta_{\mathbf{s}}(b|a)}{\left(1 - \sum_{i \geq k_n} f(i)\right)}$$

Proof Let $\mathbf{w} \in \tilde{\mathcal{T}}$ and fix $a, b \in \mathcal{A}$. Note that for $i = k_n$, by assumption we have for all $c, c' \in \mathcal{A}$

$$\left|\frac{\theta_{c\mathbf{w}}(b|a)}{\theta_{c'\mathbf{w}}(b|a)} - 1\right| \leq f(k_n). \tag{4}$$

According Lemma 2, $\tilde{\theta}_{\mathbf{w}}(b|a)$ is weighted average of $\theta_{c\mathbf{w}}(b|a)$, $c \in \mathcal{A}$. Hence, there are largest $\theta_{d\mathbf{w}}(b|a)$ and smallest $\theta_{d'\mathbf{w}}(b|a)$ such that

$$\theta_{d\mathbf{w}}(b|a) \leq \tilde{\theta}_{\mathbf{w}}(b|a) \leq \theta_{d'\mathbf{w}}(b|a). \tag{5}$$

Combining (4) with (5) and straightforward elementary algebra shows that $\forall c \in \mathcal{A}$

$$\tilde{\theta}_{\mathbf{w}}(b|a)\big(1 - f(k_n)\big) \leq \theta_{c\mathbf{w}}(b|a) \leq \big(1 + f(k_n)\big)\tilde{\theta}_{\mathbf{w}}(b|a).$$

Proceeding inductively, for all $\mathbf{s} \in \mathcal{T}_{\mathbf{w}}$ we have

$$\left(\prod_{i \geq k_n} \big(1 - f(i)\big)\right)\tilde{\theta}_{\mathbf{w}}(b|a) \leq \theta_{\mathbf{s}}(b|a) \leq \left(\prod_{i \geq k_n} \big(1 + f(i)\big)\right)\tilde{\theta}_{\mathbf{w}}(b|a).$$

Now, first part of the lemma implies that

$$\left(1 - \sum_{i \geq k_n} f(i)\right) \max_{\mathbf{s} \in \mathcal{T}_{\mathbf{w}}} \theta_{\mathbf{s}}(b|a) \leq \tilde{\theta}_{\mathbf{w}}(b|a) \leq \frac{\min_{\mathbf{s} \in \mathcal{T}_{\mathbf{w}}} \theta_{\mathbf{s}}(b|a)}{\left(1 - \sum_{i \geq k_n} f(i)\right)}.$$

□

Definition 3: For all sequences (X_1^n, Y_1^n) obtained from the channel model $p_{\mathcal{T},q}$, let $\hat{\mathcal{T}} = \mathcal{A}^{k_n}$ with $k_n = \alpha_n \log n$ for some function $\alpha_n = \mathcal{O}(1)$. For $\mathbf{s} \in \hat{\mathcal{T}}$, let $\mathbf{Y}_{\mathbf{s}}^a$ be the sequence of output symbols that follows the output string $\mathbf{s}$, and correspond to the input $x = a$. Hence, the length of $\mathbf{Y}_{\mathbf{s}}^a$ is

$$N_n(\mathbf{s}, a) = \sum_{k=1}^{n} \mathbb{1}\{\mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{k-1}) = \mathbf{s}, X_k = a\},$$

the number of occurrences of symbol b in $\mathbf{Y}_{\mathbf{s}}^a$ is $n_{\mathbf{s}}(b, a)$, where

$$n_{\mathbf{s}}(b, a) = \sum_{k=1}^{n} \mathbb{1}\{\mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{k-1}) = \mathbf{s}, Y_k = b, X_k = a\}.$$

We define the naive estimate of $\tilde{\theta}_{\mathbf{s}}(b|a)$ as

$$\hat{\theta}_{\mathbf{s}}(b|a) = \frac{n_{\mathbf{s}}(b, a)}{N_n(\mathbf{s}, a)}$$

Furthermore, let $N_n(\mathbf{s}) = \sum_{a \in \mathcal{A}} N_n(\mathbf{s}, a)$. □

Remark Note that $\mathbf{Y}_{\mathbf{s}}^a$ is *i.i.d* only if $\mathbf{s} \in \mathcal{T}$, the set of states for the true model. In general, since we do not necessarily know if any of $n_{\mathbf{s}}(b, a)$ are close to their stationary frequencies, there is no obvious reason why $\hat{\theta}_{\mathbf{s}}(b|a)$ shall reflect $\tilde{\theta}_{\mathbf{s}}(b|a)$. □

Let $\nu_j = \sum_{i \geq j} f(i)$. Note that $\nu_j \to 0$ as $j \to \infty$ and that $-\nu_j \log \nu_j \to 0$ as $\nu_j \to 0$.

Definition 4: Given a sample sequence with size n obtained from the channel model $p_{\mathcal{T},q}$, we define the set of good states, denoted by $\tilde{G}$, as

$$\tilde{G} = \big\{\mathbf{w} \in \tilde{\mathcal{T}} : \forall a \in \mathcal{A},\, N_n(\mathbf{w}, a) \geq \max\{n\nu_{k_n} \log \frac{1}{\nu_{k_n}}, |\mathcal{A}|^{k_n+1} \log^2 n\}\big\}.$$

Remark Note that a state is good if the count of the state is $\geq n^{\alpha_n} \log^2 n = 2^{k_n} \log^2 n$. Therefore, if $2^{k_n} \log^2 n \geq n$, or equivalently $k_n \geq \log n - 2 \log \log n$, no state will be good and the Theorem below becomes vacuously true. This is not a fundamental weakness in this line of argument—it is known that k_n has to scale logarithmically with n for proper estimation to hold. □

Remark Observe also that because we do not assume the source has mixed, the theorem below does not imply that the parameters are accurate for contexts shorter than k_n. This is perhaps counterintuitive at first glance, but the below result holds not because of mixing, but because of the fall-off of dependencies. $\square$

Theorem 6: Let $k_n = \alpha_n \log n$. With probability (conditioned on $Y_{-\infty}^0$) $\geq 1 - \frac{1}{2^{|\mathcal{A}|^{k_n+1}\log n}}$, for all $a \in \mathcal{A}$, $Y_{-\infty}^0$ and $\mathbf{s} \in \tilde{G}$ simultaneously

$$||\tilde{\theta}_{\mathbf{s}}(\cdot|a) - \hat{\theta}_{\mathbf{s}}(\cdot|a)||_1 \leq 2\sqrt{\frac{\ln 2}{\log n} + \frac{\ln 2}{\log \frac{1}{\nu_{k_n}}}}$$

Proof As before, let $\nu_{k_n} = \sum_{i \geq k_n} f(i)$ and let n be large enough that $\nu_{k_n} \leq \frac{1}{2}$. Note that Lemma 5 implies that for all sequences $(x_1^n, y_1^n) \in \mathcal{A}^n \times \mathcal{A}^n$ (all steps hold for all $Y_{-\infty}^0$)

$$\begin{aligned} p_{\mathcal{T},q}(y_1^n|x_1^n, Y_{-\infty}^0) &\leq \frac{1}{(1-\nu_{k_n})^n} \prod_{\mathbf{s}\in\tilde{\mathcal{T}}} \prod_{a,b\in\mathcal{A}} \tilde{\theta}_{\mathbf{s}}(b|a)^{n_{\mathbf{s}}(b,a)} \\ &\leq 4^{n\nu_{k_n}} \prod_{\mathbf{s}\in\tilde{\mathcal{T}}} \prod_{a,b\in\mathcal{A}} \tilde{\theta}_{\mathbf{s}}(b|a)^{n_{\mathbf{s}}(b,a)} \end{aligned}$$

where $\tilde{\theta}_{\mathbf{s}}(b|a)$, $n_{\mathbf{s}}(b,a)$ are defined in Definition 3 and the second inequality is because $(\frac{1}{1-t})^n \leq 4^{nt}$ whenever $0 \leq t \leq \frac{1}{2}$. Now, let B_n be the set of all sequences that satisfy

$$4^{n\nu_{k_n}} \prod_{\mathbf{s}\in\tilde{\mathcal{T}}} \prod_{a,b\in\mathcal{A}} \tilde{\theta}_{\mathbf{s}}(b|a)^{n_{\mathbf{s}}(b,a)} \leq \frac{\prod_{\mathbf{s}\in\tilde{\mathcal{T}}} \prod_{a,b\in\mathcal{A}} \hat{\theta}_{\mathbf{s}}(b|a)^{n_{\mathbf{s}}(b,a)}}{2^{2\cdot|\mathcal{A}|^{k_n+1}\log n}}.$$

Now using a construction similar to the context tree weighting algorithm [27], we obtain a distribution p_c satisfying

$$p_c(y_1^n|x_1^n, Y_{-\infty}^0) \geq \frac{\prod_{\mathbf{s}\in\tilde{\mathcal{T}}} \prod_{a,b\in\mathcal{A}} \hat{\theta}_{\mathbf{s}}(b|a)^{n_{\mathbf{s}}(b,a)}}{2^{|\mathcal{A}|^{k_n+1}\log n}}.$$

Hence, for all sequences in B_n, we have

$$\begin{aligned} p_c(y_1^n|x_1^n, Y_{-\infty}^0) &\geq \frac{\prod_{\mathbf{s}\in\tilde{\mathcal{T}}} \prod_{a,b\in\mathcal{A}} \hat{\theta}_{\mathbf{s}}(b|a)^{n_{\mathbf{s}}(b,a)}}{2^{|\mathcal{A}|^{k_n+1}\log n}} \\ &\geq \frac{4^{(n\nu_{k_n}+|\mathcal{A}|^{k_n+1}\log n)} \prod_{\mathbf{s}\in\tilde{\mathcal{T}}} \prod_{a,b\in\mathcal{A}} \tilde{\theta}_{\mathbf{s}}(b|a)^{n_{\mathbf{s}}(b,a)}}{2^{|\mathcal{A}|^{k_n+1}\log n}} \\ &\geq p_{\mathcal{T},q}(y_1^n|x_1^n, Y_{-\infty}^0) 2^{|\mathcal{A}|^{k_n+1}\log n}. \end{aligned}$$

Thus, B_n is the set of sequences y_1^n such that p_c assigns a much higher probability than $p_{\mathcal{T},q}$. Such a set B_n can not have high probability under $p_{\mathcal{T},q}$.

$$\begin{aligned} p_{\mathcal{T},q}(B_n) &= p_{\mathcal{T},q}\left\{ y_1^n : p_c(y_1^n|x_1^n, Y_{-\infty}^0) \geq p_{\mathcal{T},q}(y_1^n|x_1^n, Y_{-\infty}^0) 2^{|\mathcal{A}|^{k_n+1}\log n} \right\} \\ &\leq \sum_{y_1^n\in B_n} p_c(y_1^n|x_1^n, Y_{-\infty}^0) 2^{-|\mathcal{A}|^{k_n+1}\log n} \leq 2^{-|\mathcal{A}|^{k_n+1}\log n}. \end{aligned}$$

Therefore, with probability $\geq 1 - 2^{-|\mathcal{A}|^{k_n+1}\log n}$, we have

$$\prod_{\mathbf{s}\in\tilde{\mathcal{T}}} \prod_{a,b\in\mathcal{A}} \tilde{\theta}_{\mathbf{s}}(b|a)^{n_{\mathbf{s}}(b,a)} \geq \frac{\prod_{\mathbf{s}\in\tilde{\mathcal{T}}} \prod_{a,b\in\mathcal{A}} \hat{\theta}_{\mathbf{s}}(b|a)^{n_{\mathbf{s}}(b,a)}}{4^{(|\mathcal{A}|^{k_n+1}\log n + n\nu_{k_n})}}$$

which implies simultaneously for all $\mathbf{s} \in \tilde{\mathcal{T}}$ and for all $a \in \mathcal{A}$

$$\prod_{b\in\mathcal{A}} \tilde{\theta}_{\mathbf{s}}(b|a)^{n_{\mathbf{s}}(b,a)} \geq \frac{\prod_{b\in\mathcal{A}} \hat{\theta}_{\mathbf{s}}(b|a)^{n_{\mathbf{s}}(b,a)}}{4^{(|\mathcal{A}|^{k_n+1}\log n + n\nu_{k_n})}}.$$

The above equation implies that $\tilde{\theta}_{\mathbf{s}}(\cdot|a)$ and $\hat{\theta}_{\mathbf{s}}(\cdot|a)$ are close distributions, since we can rearrange (take logarithm and divide both sides by $N_n(\mathbf{s},a)$) to obtain

$$\sum_{b\in\mathcal{A}} \frac{n_{\mathbf{s}}(b,a)}{N_n(\mathbf{s},a)} \log \frac{\hat{\theta}_{\mathbf{s}}(b|a)}{\tilde{\theta}_{\mathbf{s}}(b|a)} = D\Big(\hat{\theta}_{\mathbf{s}}(\cdot|a) \| \tilde{\theta}_{\mathbf{s}}(\cdot|a)\Big) \leq \frac{2(|\mathcal{A}|^{k_n+1}\log n + n\nu_{k_n})}{N_n(\mathbf{s},a)},$$

where the first equality follows by writing out the value of the naive estimate, $\hat{\theta}_{\mathbf{s}}(b|a) = n_{\mathbf{s}}(b,a)/N_n(\mathbf{s},a)$. Since (see for example [32])

$$\frac{1}{2\ln 2} ||\hat{\theta}_{\mathbf{s}}(\cdot|a) - \tilde{\theta}_{\mathbf{s}}(\cdot|a)||_1^2 \leq D\Big(\hat{\theta}_{\mathbf{s}}(\cdot|a) \| \tilde{\theta}_{\mathbf{s}}(\cdot|a)\Big),$$

for all $\mathbf{s} \in \tilde{\mathcal{T}}$ and $a \in \mathcal{A}$, we now have with confidence bigger than $1 - 2^{-|\mathcal{A}|^{k_n+1}\log n}$ that

$$||\tilde{\theta}_{\mathbf{s}}(\cdot|a) - \hat{\theta}_{\mathbf{s}}(\cdot|a)||_1 \leq \sqrt{\frac{(\ln 2)(|\mathcal{A}|^{k_n+1}\log n + n\nu_{k_n})}{N_n(\mathbf{s},a)}}.$$

The Theorem follows from our Definition 4 of good states. □

When the dependencies among strings die down exponentially, we can strengthen Theorem 6 to get convergence rate polynomial in n.

Theorem 7: Suppose $f(i) = \gamma^i$ for some $0 < \gamma < 1$. Let ζ be a nonnegative constant such that $\zeta \geq \frac{-\log\gamma}{\log\mathcal{A} - \log\gamma}$. For $k_n = \frac{\log n}{\log\mathcal{A} - \log\gamma}$ and define

$$\tilde{G} \triangleq \{\mathbf{w} \in \tilde{\mathcal{T}} : \forall a \in \mathcal{A}, \quad N_n(\mathbf{w},a) \geq n^{\zeta + \frac{\log\mathcal{A}}{\log\mathcal{A} - \log\gamma}}\}.$$

Then, for all $\mathbf{s} \in \tilde{G}$ with probability greater than $1 - 2^{-|\mathcal{A}|^{k_n+1}\log n}$ simultaneously

$$||\tilde{\theta}_{\mathbf{s}}(\cdot|a) - \hat{\theta}_{\mathbf{s}}(\cdot|a)||_1 \leq 2\sqrt{\frac{\ln 2 \cdot \big((1-\gamma)|\mathcal{A}|\log n + 1\big)}{(1-\gamma)n^{\zeta}}}.$$ □

Proof The proof is similar to Theorem 6, but involves more careful but elementary algebra specific to the exponential decay case.

Remark According to definition of good states in Theorem 7 and the fact that $d(\tilde{\mathcal{T}}) = k_n$, we obtain

$$|\tilde{G}| \leq n^{-\frac{\log\gamma}{\log|\mathcal{A}| - \log\gamma}}, |\tilde{\mathcal{T}}| = n^{\frac{\log|\mathcal{A}|}{\log|\mathcal{A}| - \log\gamma}}$$

implying that if $\gamma \leq 1/|\mathcal{A}|$, all states of $\tilde{\mathcal{T}}$ can potentially be good. □

Note that the parameters $\tilde{\theta}$ associated with any good state can be well estimated from the sample while the rest may not be accurate. From Example 3, we know that the stationary probabilities may be a very sensitive function of the parameters associated with states. It is therefore perfectly possible that we estimate the parameters at all states reasonably well, but are unable to gauge what the stationary probabilities of any state may be. How do we tell, therefore, if we can trust our naive counts of states?

To find deviation bounds for stationary distribution of good states, we construct a new process $\{Z_m\}_{m=1}^{\infty}$, $Z_m \in \mathcal{T}$ from the process $\{Y_n\}_{n=1}^{\infty}$. If Y_{n_m} is the $(m+1)^{th}$ symbol in the sequence $\{Y_n\}_{n=1}^{\infty}$ such that $\mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{n_m}) \in \tilde{G}$, then $Z_m = \mathbf{c}_{\mathcal{T}}(Y_{-\infty}^{n_m})$. The strong Markov property allows us to characterize $\{Z_m\}_{m=1}^{\infty}$ as a Markov process with transitions that are lower bounded by those transitions of the process $\{Y_n\}_{n=1}^{\infty}$ that can be well estimated by the Theorems above. More specifically, let $T_0 = \min\{j \geq 0 : \mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{j}) \in \tilde{G}\}$

and let $Z_0 = \mathbf{c}_{\mathcal{T}}(Y_{-\infty}^{T_0})$. For all $m \geq 1$, T_m is the $(m+1)'$th occurrence of a good state in the sequence $\{Y_n\}_{n=1}^{\infty}$, namely

$$T_m = \min\{j \geq T_{m-1} : \mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{j}) \in \tilde{G}\},$$

and $Z_m = \mathbf{c}_{\mathcal{T}}(Y_{-\infty}^{T_m})$. Note that T_m is a *stopping time* [8], and therefore $\{Z_m\}_{m=1}^{\infty}$ is a Markov chain by itself. Let $\tilde{B} = \{\mathbf{s} \in \tilde{\mathcal{T}} : \mathbf{s} \notin \tilde{G}\}$. The transitions between states $\mathbf{s}, \mathbf{s}' \in \tilde{G}$ are then the minimal, non-negative solution of the following set of equations in $\{Q(\mathbf{s}|\mathbf{s}'') : \mathbf{s}'' \in \mathcal{A}^{k_n}, \mathbf{s} \in \tilde{G}\}$

$$Q(\mathbf{s}|\mathbf{s}') = p_{\mathcal{T},q}\big(\mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{1}) = \mathbf{s}|\mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{0}) = \mathbf{s}'\big) + \sum_{\mathbf{s}'' \in \tilde{B}} p_{\mathcal{T},q}\big(\mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{1}) = \mathbf{s}''|\mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{0}) = \mathbf{s}'\big)Q(\mathbf{s}|\mathbf{s}'')$$

An important point to note here is that if $\mathbf{s}$ and $\mathbf{s}'$ are good states,

$$Q(\mathbf{s}|\mathbf{s}') \geq p_{\mathcal{T},q}\big(\mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{1}) = \mathbf{s}|\mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{0}) = \mathbf{s}'\big),$$

and the lower bound above can be well estimated from the sample as shown in Theorem 6.

Property 1: A few properties about $\{Z_m\}$ are in order. $\{Z_m\}$ is constructed from an irreducible process $\{Y_n\}$, thus $\{Z_m\}$ is irreducible as well. Since $\{Y_n\}$ is positive recurrent, so is $\{Z_m\}$. But despite $\{Y_n\}$ being aperiodic, $\{Z_n\}$ could be periodic as in the Example below. But periodicity of $\{Z_m\}$ can be determined by $\tilde{G}$ alone (because $\mathcal{T}$, while unknown, is a full, finite $\mathcal{A}$-ary tree). □

Example 6: Let $\{Y_n\}$ be a process generated by context tree model $p_{\mathcal{T},q}$ with $\mathcal{T} = \{11, 01, 10, 00\}$ and $q(1|11) = \frac{1}{2}, q(1|01) = \epsilon, q(1|10) = 1-\epsilon, q(1|00) = \frac{1}{2}$. If $\epsilon > 0$, then $p_{\mathcal{T},q}$ represents a stationary aperiodic Markov process. If $\{Z_n\}$ be the restriction of process $\{Y_n\}$ to $\tilde{G} = \{01, 10\}$, the restricted process will be periodic with period 2. □

Property 2: Suppose $\{Z_m\}$ is aperiodic. Let μ_Y and μ_Z denote the stationary distribution of the processes $\{Y_n\}$ and $\{Z_m\}$, respectively, with$n$ samples of a sequence $\{Y_n\}$ yielding m_n samples of $\{Z_m\}$. Similarly, let $\mu_Z(\mathbf{s})$denote the stationary probability of the event $\mathbf{s} \preceq Z$. Then for all $\mathbf{s}, \mathbf{s}' \in \tilde{G}$ with $\mu_Y(\mathbf{s}') \neq 0$,

$$\frac{\mu_Y(\mathbf{s})}{\mu_Y(\mathbf{s}')} \overset{wp1}{=} \frac{\lim_{n\to\infty} \sum_{i=1}^{n} \frac{\mathbb{1}(\mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{i})=\mathbf{s})}{n}}{\lim_{n\to\infty} \sum_{i=1}^{n} \frac{\mathbb{1}(\mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{i})=\mathbf{s}')}{n}} = \frac{\lim_{m_n\to\infty} \sum_{i=1}^{m} \frac{\mathbb{1}(\mathbf{s}\preceq Z_i)}{m}}{\lim_{m_n\to\infty} \sum_{i=1}^{m} \frac{\mathbb{1}(\mathbf{s}'\preceq Z_i)}{m}} \overset{wp1}{=} \frac{\mu_Z(\mathbf{s})}{\mu_Z(\mathbf{s}')}. \qquad \square$$

For any (good) state $\mathbf{s}$, let $G_{\mathbf{s}} \subset \mathcal{A}$ be the set of letters that take $\mathbf{s}$ to another good state,

$$G_{\mathbf{s}} = \{b \in \mathcal{A} : \mathbf{c}_{\tilde{\mathcal{T}}}(\mathbf{s}b) \in \tilde{G}\}. \tag{6}$$

Our confidence in the empirical counts of good states matching their (aggregated) stationary probabilities follows from a coupling argument, and depends on the following parameter

$$\eta_{\tilde{G}} = \min_{\mathbf{u},\mathbf{v} \in \tilde{G}} \sum_{b \in G_{\mathbf{u}} \cap G_{\mathbf{v}}} \min\{\tilde{q}(b|\mathbf{u}), \tilde{q}(b|\mathbf{v})\}. \tag{7}$$

Note that for the channel model $p_{\mathcal{T},q}$, we have

$$\tilde{\theta}(b|\mathbf{u}) = \sum_{a\in\mathcal{A}} P(X_1 = a)P(Y_1 = b|\mathbf{c}_{\tilde{\mathcal{T}}}(Y_{-\infty}^{0}) = \mathbf{u}, X_1 = a) = \sum_{a\in\mathcal{A}} p_a \theta_{\mathbf{u}}(b|a).$$

Again, Theorems 6 and 7 allow us to estimate $\eta_{\tilde{G}}$ from the sample since it only depends on parameters associated with good states. The counts of various $\mathbf{w} \in \tilde{G}$ now concentrates as shown in the Theorem below, and how good the concentration is can be estimated as a function of $\eta_{\tilde{G}}$ (and ν_{k_n}) and the total count of all states in $\tilde{G}$ as below. Now $\tilde{G}$ as well as $\eta_{\tilde{G}}$ are well estimated from the sample—thus we can look at the data to interpret the empirical counts of various substrings of the data. Let $\Phi_j = \sum_{i\geq j} \nu_i$. For the following theorem, we require ν_i to be summable. Thus, Φ_j is finite for all j and decreases to 0 as j increases. If $f(i) \sim \gamma^i$, then Φ_j also diminishes as γ^j. But $f(i) \sim \frac{1}{i^r}$ diminishes polynomially, then

Φ_j diminishes as $1/j^{r-2}$. If $f(i) = 1/i^{2+\eta}$ for any $\eta > 0$, we therefore satisfy the summability of ν_i. However, $f(i)$ can also diminish as $1/(i^2 \operatorname{poly}(\log i))$ for appropriate polynomials of $\log i$ for the counts of good states to converge.

Theorem 8: If $\{Z_m\}_{m=1}^{\infty}$ is aperiodic, then for any $t > 0$, $Y_{-\infty}^0$ and $\mathbf{w} \in \tilde{G}$ we have

$$p_{\mathcal{T},q}(|N_n(\mathbf{w}) - \tilde{n}\frac{\mu(\mathbf{w})}{\mu(\tilde{G})}| \geq t | Y_{-\infty}^0) \leq 2\exp\left(-\frac{t^2}{2\tilde{n}}\left(\frac{\eta_{\tilde{G}}^{k_n}(1-\Phi_{k_n})}{4\ell_n + \eta_{\tilde{G}}^{k_n}(1-\Phi_{k_n})}\right)^2\right)$$

where ℓ_n is the smallest integer such that $\Phi_{\ell_n} \leq \frac{1}{n}$, $\tilde{n}$ is the total count of good states in the sample and μ is the stationary distribution of $p_{\mathcal{T},q}$.

Proof We define

$$V_i = \mathbb{E}[N_n(\mathbf{w})|Z_0 Z_1, \ldots, Z_i],$$

and observe that $\{V_i\}_{i=1}^{\tilde{n}}$ is a Doob Martingale. Note that $V_0 = \mathbb{E}[N_n(\mathbf{w})|Z_0]$ and $V_{\tilde{n}} = N_n(\mathbf{w})$.

Remark To summarize, we first bound the differences $|V_{i+1} - V_i|$ of the martingale using a coupling argument on two copies of the chain $\{Z_n\}$. Since the memory of the process $p_{\mathcal{T},q}$ could be large, in our proof the coupled chains never actually coalesce in the usual sense but enough that the chance they diverge again within n samples is less than $1/n$. This is where the parameter ℓ_n comes in as well. Once we bound the differences in the martingale $\{V_i\}_{i=1}^{\tilde{n}}$, the theorem follows as an easy application of Azuma's inequality. $\square$

Now since for all $i \geq 0$

$$\begin{aligned}|V_i - V_{i-1}| &= |\mathbb{E}[N_n(\mathbf{w})|Z_0, \ldots, Z_i] - \mathbb{E}[N_n(\mathbf{w})|Z_0, \ldots, Z_{i-1}]| \\ &\leq \max_{Z_i^{'}, Z_i^{''}} \left|\mathbb{E}[N_n(\mathbf{w})|Z_0, \ldots, Z_i^{'}] - \mathbb{E}[N_n(\mathbf{w})|Z_0, \ldots, Z_i^{''}]\right|,\end{aligned}$$

we bound the maximum change in $N_n(\mathbf{w})$ if the i'th good state was changed into another (good) state. To do so, we use a coupling argument as follows. Let $\tilde{G}$ be the set of good states from Definition 4, and suppose good states occur $\tilde{n}$ times in the sequence. Suppose there are sequences $\{Z_i^{'}\}$ (starting from state $Z_i^{''}$) and $\{Z_i^{''}\}$ (starting from state $Z_i^{'}$), both faithful copies of $\{Z_i\}$ yet coupled with a joint distribution ω to be described below. From the coupling argument of Section IV-B, we have for $\mathbf{w} \in \tilde{G}$ (hence $|\mathbf{w}| = k_n$) for all ω

$$|\mathbb{E}[N_n(\mathbf{w})|Z_0, \ldots, Z_i^{'}] - \mathbb{E}[N_n(\mathbf{w})|Z_0, \ldots, Z_i^{''}]| \leq \sum_{j=i+1}^{\tilde{n}} \omega(Z_j^{'} \stackrel{k_n}{\not\approx} Z_j^{''}), \tag{8}$$

where $Z_j^{'} \stackrel{k_n}{\approx} Z_j^{''}$ if $\mathbf{c}_{\mathcal{A}^{k_n}}(Z_j^{'}) = \mathbf{c}_{\mathcal{A}^{k_n}}(Z_j^{''})$.

a) Description of Coupling: Suppose we have $\{Z_i^{'}\}_{i=1}^{j}$ and $\{Z_i^{''}\}_{i=1}^{j}$. To obtain $Z_{j+1}^{'}$ and $Z_{j+1}^{''}$, starting from states $Z_j^{'}$ and $Z_j^{''}$ we run copies $\{Y_{ji}^{'}\}_{i\geq 1}$ and $\{Y_{ji}^{''}\}_{i\geq 1}$ of coupled chains individually faithful to $p_{\mathcal{T},q}$. Then $Z_j^{'}$ is the state corresponding to the first time $\{Z_j, Y_{ji}^{'}\}_{i\geq 1}$ hits a context in $\tilde{G}$. Similarly for $Z_j^{''}$. Specifically, the chains $\{Y_{ji}^{'}\}_{i\geq 1}$ and $\{Y_{ji}^{''}\}_{i\geq 1}$ are coupled as follows. We generate a number U_{j1} uniformly distributed $\in [0,1]$. Given ($Z_j^{'}$ and $Z_j^{''}$) with suffixes $\mathbf{u}$ and $\mathbf{v}$ respectively in $\tilde{G}$, we let $G_{\mathbf{u}} \in \mathcal{A}$ (and $G_{\mathbf{v}}$ similarly) be the set of symbols in $\mathcal{A}$ defined as in (6). We split the interval from 0 to 1 as follows: for all $a \in \mathcal{A}$, we assign intervals $r(a)$ of length $\min\{q(a|\mathbf{u}), q(a|\mathbf{v})\}$, in the following order: we first stack the above intervals corresponding to $a \in G_{\mathbf{u}} \cap G_{\mathbf{v}}$ (in any order) starting from 0, and then we put in the intervals corresponding to all other symbols. Now let,

$$(Y_{j1}^{'}, Y_{j1}^{''}) = (a, a) \text{ if } U_{j1} \in r(a).$$

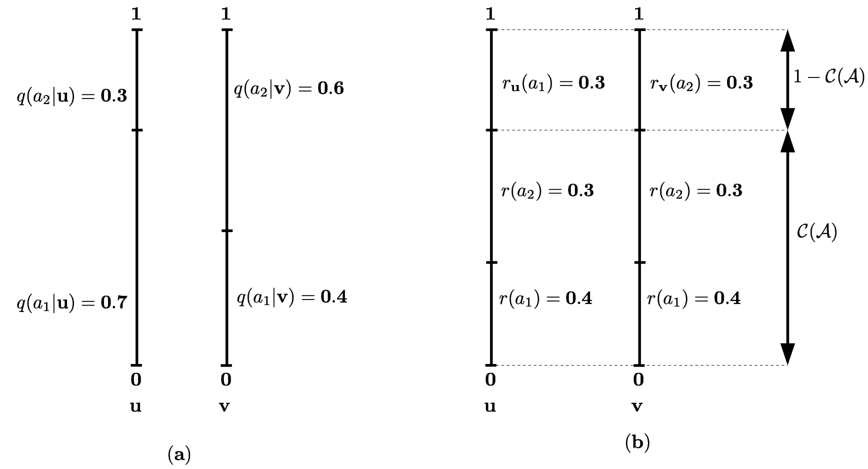

Fig. 4. (a) The conditional probabilities with which $Y_{j1}^{'}$ and $Y_{j1}^{''}$ have to be chosen respectively are $q(\cdot|\mathbf{u})$ and $q(\cdot|\mathbf{v})$. The line on the left determines the choice of $Y_{j1}^{'}$ and the one on the right the choice of $Y_{j1}^{''}$. For example, if U_{j1} is chosen uniformly in [0,1], the probability of choosing $Y_{j1}^{'} = a_1$ is $q(a_1|\mathbf{u})$. Instead of choosing $Y_{j1}^{'}$ and $Y_{j1}^{''}$ independently, we will reorganize the intervals in the lines so as to encourage $Y_{j1}^{'} = Y_{j1}^{''}$. (b) Reorganizing the interval $[0, 1]$ according to the described construction. Here $r(a_1) = \min\{q(a_1|\mathbf{u}), q(a_1|\mathbf{v})\}$ and similarly for $r(a_2)$. If U_{j1} falls in the interval corrsponding to $r(a_1)$, then $(Y_{j1}^{'}, Y_{j1}^{''}) = (a_1, a_1)$. If $U_{j1} > C(\mathcal{A})$ in this example, then $(Y_{j1}^{'}, Y_{j1}^{''}) = (a_1, a_2)$. When U_{j1} is chosen uniformly in [0,1], the probability $Y_{j1}^{'}$ outputs any symbol is the same as in the picture on the left, similarly for $Y_{j2}^{'}$.

Let

$$C(\mathcal{A}) = \sum_{a\in\mathcal{A}} r(a) = \sum_{a\in\mathcal{A}} \min\{q(a|Z_j^{'}), q(a|Z_j^{''})\}. \tag{9}$$

be the part of the interval is already filled up. Thus if $U_{j1} < C(\mathcal{A})$, equivalently with probability $C(\mathcal{A})$, the two chains output the same symbol. We use the rest of the interval $[C(A), 1]$ in any valid way to satisfy the fact that $Y_{j1}^{'}$ is distributed as $p_{\mathcal{T},q}(\cdot|Z_j^{'})$ and $Y_{j1}^{''}$ is distributed as $p_{\mathcal{T},q}(\cdot|Z_j^{''})$. For one standard approach, for all a assign

$$r_{\mathbf{u}}(a) = (q(a|\mathbf{u}) - q(a|\mathbf{v}))^{+} = \max\{q(a|\mathbf{u}) - q(a|\mathbf{v}), 0\}$$

and similarly $r_{\mathbf{v}}(a)$. Note that only one of $r_{\mathbf{u}}(a)$ and $r_{\mathbf{v}}(a)$ can be strictly positive and that for all a, $r(a) + r_{\mathbf{u}}(a) = q(a|\mathbf{u})$ while $r(a) + r_{\mathbf{v}}(a) = q(a|\mathbf{v})$. Therefore,

$$\sum_{a\in\mathcal{A}} r_{\mathbf{u}}(a) = \sum_{a\in\mathcal{A}} r_{\mathbf{v}}(a) = 1 - C(\mathcal{A}).$$

An example of such construction for binary alphabet is illustrated in Fig. 4. in which we have assumed $G_{\mathbf{u}} \cap G_{\mathbf{v}} = \{a_1\}$. We will keep two copies of the interval $[C(\mathcal{A}), 1]$, and if $U_{j1} > C(\mathcal{A})$ we output $(Y_{j1}^{'}, Y_{j1}^{''})$ based on where U_{j1} falls in both copies. We will stack the first copy of $[C(\mathcal{A}), 1]$ with intervals of length $r_{\mathbf{u}}(a)$ for all a and the second copy of $[C(\mathcal{A}), 1]$ with intervals length $r_{\mathbf{v}}(a)$ for all a. We say $U_{j1} \in (r_{\mathbf{u}}(a), r_{\mathbf{v}}(b))$ if $U_{j1} \in r_{\mathbf{u}}(a)$ in the first copy and $U_{j1} \in r_{\mathbf{v}}(b)$ in the second copy. Furthermore,

$$(Y_{j1}^{'}, Y_{j1}^{''}) = (a, b) \text{ if } U_{j1} \in (r_{\mathbf{u}}(a), r_{\mathbf{v}}(b)).$$

1) If $\mathbf{c}_{\tilde{\mathcal{T}}}(Z_j^{'} Y_{j1}^{'}) \in \tilde{G}$ and $\mathbf{c}_{\tilde{\mathcal{T}}}(Z_j^{''} Y_{j1}^{''}) \in \tilde{G}$, we have $Z_{j+1}^{'}$ and $Z_{j+1}^{''}$.

2) If $Y^{'}_{j1} = Y^{''}_{j1}$ but only one of $\mathbf{c}_{\tilde{\mathcal{T}}}(Z^{'}_j Y^{'}_{j1}) \in \tilde{G}$ and $\mathbf{c}_{\tilde{\mathcal{T}}}(Z^{''}_j Y^{''}_{j1}) \in \tilde{G}$, then we have one of $Z^{'}_{j+1}$ and $Z^{''}_{j+1}$. To get the other, we continue (according to transitions defined by $p_{\mathcal{T},q}$) only its corresponding chain till we get a good state.
3) If $Y^{'}_{j1} = Y^{''}_{j1}$, $\mathbf{c}_{\tilde{\mathcal{T}}}(Z^{'}_j Y^{'}_{j1}) \notin \tilde{G}$ and $\mathbf{c}_{\tilde{\mathcal{T}}}(Z^{''}_j Y^{''}_{j1}) \notin \tilde{G}$, we need to continue both chains. We generate $Y^{'}_{j2}, Y^{''}_{j2}$ as we did for the first samples—by generating a new random number U_{j2} uniform in $[0,1]$, and by coupling as in (9) the distributions $q(\cdot|Z^{'}_j Y^{'}_{j1})$ and $q(\cdot|Z^{''}_j Y^{''}_{j1})$ respectively. And continue in this fashion so long as the samples in the two chains remain equal but do not hit good contexts. This will be case that will be most important for us later on.
4) If $Y^{'}_{jl} \neq Y^{''}_{jl}$ at any point and neither chain has seen a good state yet, we just run the chains independently from that point on for how long it takes each to hit a good aggregated state.

b) Analysis of coupling: For any r, let $Z^{'}_r \sim Z^{''}_r$ denote the following event that is a subset of case (1) in the list above,

$$\Big\{Y^{'}_{r1} = Y^{''}_{r1} \text{ and } \mathbf{c}_{\tilde{\mathcal{T}}}(Z^{'}_{r-1} Y^{'}_{r1}) \in \tilde{G} \text{ and } \mathbf{c}_{\tilde{\mathcal{T}}}(Z^{''}_{r-1} Y^{''}_{r1}) \in \tilde{G}\Big\}.$$

From (7) and using Lemma 5, we can easily show

$$\omega(Z^{'}_r \sim Z^{''}_r | Z^{'}_{r-1}, Z^{''}_{r-1}) \geq \eta_{\tilde{G}}(1 - \nu_{k_n}), \tag{10}$$

where the $1-\nu_{k_n}$ term comes because the parameter $\eta_{\tilde{G}}$ is defined on the aggregated parameters, but $Y^{'}_j$ and $Y^{''}_j$ evolve according to $p_{\mathcal{T},q}$. Furthermore, if $Z^{'}_i \sim Z^{''}_i$ for the k_n consecutive samples $j - k_n + 1 \leq i \leq j$, then we have $Z^{'}_j \overset{k_n}{\approx} Z^{''}_j$. To proceed, once $Z^{'}_j \overset{k_n}{\approx} Z^{''}_j$, we would like them to coalesce tighter in every subsequent step, namely we want for all $1 \leq l \leq n$, $Z^{'}_{j+l} \overset{k_n+l}{\approx} Z^{''}_{j+l}$. Starting from $Z^{'}_j \overset{k_n}{\approx} Z^{''}_j$, one way we can have $Z^{'}_{j+1} \overset{k_n+1}{\approx} Z^{''}_{j+1}$ is if $Z^{'}_{j+1} \sim Z^{''}_{j+1}$, or if the chains $\{Y^{'}_{ji}\}_{i\geq 1}$ and $\{Y^{''}_{ji}\}_{i\geq 1}$ evolve through a sequence of $m > 1$ steps before hitting a context in $\tilde{G}$ on the m'th step with $Y^{'}_{jl} = Y^{''}_{jl}$ for each $l \leq m$. This is the situation in case 3 of the list above, but in addition in each step l,

$$\mathbf{c}_{\mathcal{A}^{k_n}}(\{Z_j, Y^{'}_{ji}\}_{i=1}^{l}) = \mathbf{c}_{\mathcal{A}^{k_n}}(\{Z_j, Y^{'}_{ji}\}_{i=1}^{l}),$$

since $Z^{'}_j \overset{k_n}{\approx} Z^{''}_j$. Therefore, both chains will hit a common good context in $\tilde{G}$ in m steps. But in addition we will also have

$$Z^{'}_{j+1} \overset{k_n+m}{\approx} Z^{''}_{j+1}.$$

Because of the way we have set up our coupling, the probability

$$\begin{aligned}
\omega(Y^{'}_{j1} = Y^{''}_{j1} | Z^{'}_j \overset{k_n}{\approx} Z^{''}_j) &= \sum_{a\in\mathcal{A}} \min\Big\{q(a|Z^{'}_j), q(a|Z^{''}_j)\Big\} \\
&\geq \sum_{a\in\mathcal{A}} \tilde{q}\Big(a|\mathbf{c}_{\tilde{\mathcal{T}}}(Z^{'}_j)\Big)(1-\nu_{k_n}) \\
&= 1 - \nu_{k_n},
\end{aligned}$$

where q and $\tilde{q}$ are the model parameters associated with $p_{\mathcal{T},q}$ and $p_{\tilde{\mathcal{T}},\tilde{q}}$ respectively. Similarly

$$\omega\left(Y^{'}_{j(l+1)} = Y^{''}_{j(l+1)} \,\middle|\, Z^{'}_j \overset{k_n}{\approx} Z^{''}_j, \{Y^{'}_{ji}\}_{i=1}^{l} = \{Y^{''}_{ji}\}_{i=1}^{l}\right) \geq 1 - \nu_{k_n+l}.$$

Therefore (no matter what m is),

$$\omega(Z^{'}_j \overset{k_n+1}{\approx} Z^{''}_j | Z^{'}_{j-1} \overset{k_n}{\approx} Z^{''}_{j-1}) \geq \prod_{l=k_n}^{\infty}(1-\nu_l) \geq 1 - \Phi_{k_n}.$$

Note that we use a very similar argument to obtain for all l

$$\omega(\exists\, \ell' \leq \ell \text{ s.t. } Z'_{j+\ell'} \overset{k_n+\ell}{\approx} Z''_{j+\ell'} | Z'_j \overset{k_n}{\approx} Z''_j) \geq \prod_{l=k_n}^{\infty} (1-\nu_l) \geq 1 - \Phi_{k_n}.$$

because the above event, $\{\exists\, \ell' \leq \ell \text{ s.t. } Z'_{j+\ell'} \overset{k_n+\ell}{\approx} Z''_{j+\ell'}\}$ can happen by going through a sequence tighter and tighter coalesced transitions of $p_{\mathcal{T},q}$ (no matter in how many steps we saw contexts in $\tilde{G}$). And we can easily strengthen the above to say for all l,

$$\omega(Z'_{j+\ell} \overset{k_n+\ell}{\approx} Z''_{j+\ell} | Z'_j \overset{k_n}{\approx} Z''_j) \geq 1 - \Phi_{k_n} \tag{11}$$

for the same reason. Indeed, we can further strengthen the above statement to note that after we see $Z'_j \overset{\ell}{\approx} Z''_j$ for any l, the chance of ever diverging is

$$\omega(\exists l > 0 \text{ s.t. } Z'_{j+l} \overset{\ell}{\not\approx} Z''_{j+l} | Z'_j \overset{\ell}{\approx} Z''_j) \leq \Phi_{\ell}. \tag{12}$$

c) Bound on Martingale differences: Let ℓ_n be as defined in the statement of this Theorem. In an abuse of notation, we say the chains $\{Z'_i\}$ and $\{Z''_i\}$ have *merged* if for any j, $Z'_j \overset{\ell_n}{\approx} Z''_j$, and let τ be the smallest number such that $Z'_\tau \overset{\ell_n}{\approx} Z''_\tau$. The probability $\tau > t\ell_n$ can be upper bounded by observing splitting the first $t\ell_n$ samples in blocks of length ℓ_n, and observing that the probability the chains merge in any single block is, using (10) k_n times and then (11) $\geq \eta_{\tilde{G}}^{k_n}(1-\Phi_{k_n})(1-\nu_{k_n})^{k_n} \geq \eta_{\tilde{G}}^{k_n}(1-\Phi_{k_n})/4$. Thus,

$$\omega(\tau > t\ell_n) \leq \Big(1 - \eta_{\tilde{G}}^{k_n}(1-\Phi_{k_n})/4\Big)^t.$$

Furthermore, $\mathbb{E}\tau$ is less than the expected number of blocks before the chains merge in any single block, thus,

$$\mathbb{E}\tau \leq \frac{4\ell_n}{\eta_{\tilde{G}}^{k_n}(1-\Phi_{k_n})}.$$

Furthermore, note that for all $j \geq i+1$,

$$\begin{aligned}\omega(Z'_j \overset{\ell_n}{\not\approx} Z''_j) &= \omega(Z'_j \overset{\ell_n}{\not\approx} Z''_j \text{ and } \tau < j) + \omega(Z'_j \overset{\ell_n}{\not\approx} Z''_j \text{ and } \tau > j)\\ &\overset{(a)}{\leq} \Phi_{\ell_n} + \omega(Z'_j \overset{\ell_n}{\not\approx} Z''_j \text{ and } \tau > j) \leq \frac{1}{n} + \omega(\tau > j).\end{aligned}$$

where inequality (a) above follows because $Z'_\tau \overset{\ell_n}{\approx} Z''_\tau$ by definition and from (12). Finally we upper bound (8)

$$\sum_{j=i+1}^{\tilde{n}} \omega(Z'_j \overset{k_n}{\not\approx} Z''_j) \leq \tilde{n} \cdot \frac{1}{n} + \sum_{j=i+1}^{\tilde{n}} \omega(\tau > j) \leq 1 + \mathbb{E}\tau \leq 1 + \frac{4\ell_n}{\eta_{\tilde{G}}^{k_n}(1-\Phi_{k_n})}.$$

The final step of the proof comes by bounding the value of $V_0 = \mathbb{E}[N_n(\mathbf{w})|Z_0]$ by a coupling argument very similar to the one above. Suppose $\{Z'_n\}$ and $\{Z''_n\}$ are coupled copies of $\{Z_n\}$, where $\{Z'_n\}$ starts from state Z_0, while $\{Z''_n\}$ starts from a state chosen randomly according to the stationary distribution of $\{Z_n\}$. The same analysis holds, and from Property 2

$$\left| V_0 - \frac{\mu(\mathbf{w})}{\mu(\tilde{G})} \right| \leq 1 + \frac{4\ell_n}{\eta_{\tilde{G}}^{k_n}(1-\Phi_{k_n})}$$

as well. (The theorem only has at most constant confidence if the upper bound above $\geq t/2\sqrt{\tilde{n}}$.) The Theorem now follows by a direct application of Azuma's inequality. $\square$

Remark Note that if the dependencies die down exponentially, namely $f(i) = \gamma^i$ for some $0 < \gamma < 1$, then $\ell_n = \lceil \log\left(n/(1-\gamma)^2\right)/\log 1/\gamma \rceil$. If the dependencies die down polynomially, namely $f(i) = 1/i^r$ for some $r > 2$, then $\ell_n \geq 2 + \left(\frac{n}{(r-1)(r-2)}\right)^{\frac{1}{r-2}}$. Furthermore, for $k_n = \mathcal{O}(\log n)$, if $\eta_{\tilde{G}}^{2k_n} \leq 1/\tilde{n}$, or $\ell_n \geq \sqrt{n}$, the theorem becomes vacuous. $\square$

VII. Conclusions

We have shown how to use data generated by potentially slow mixing Markov sources (or channels) to identify those states for which naive approaches will estimate both parameters and functions related to stationary probabilities. To do so, we require that the underlying Markov source (or channel) be well behaved—in the sense that dependencies cannot be completely arbitrary, but die down eventually. In such cases, we show that even while the source may not have mixed (explored the state space properly), certain properties can related to contexts $\mathbf{w}$, namely $\tilde{\theta}_{\mathbf{w}}$ or $\mu(\mathbf{w})/\mu(\tilde{G})$ be well estimated, if $|\mathbf{w}|$ grows as $\Theta(\log n)$. Surprisingly, we saw that it is quite possible that estimates related to contexts $\mathbf{w}$ may be good, even when estimates for suffixes of $\mathbf{w}$ fail. The reason is that Theorem 6 depends not on the source mixing, but on the dependencies dying off. This work also uncovers a lot of open problems. The above results are sufficient to say that some estimates are approximately accurate with high confidence. A natural, but perhaps difficult, question is whether we can give necessary conditions on how the data must look for any estimate to be accurate. Secondly, since we can control the evolution of the channel (or source), how—if possible—do we force the channel to explore other states better? Namely, if the channel had feedback, how could we adapt the source to better guage channel properties? Practically speaking, how do we best leverage these results in operating channels with long memory?

Acknowledgment

This work was supported by NSF Grants CCF-1065632, CCF-1018984 and EECS-1029081. The authors also thank A. Kavčić for helpful discussions.

References

[1] B. Fittingoff, "Universal methods of coding for the case of unknown statistics," in *Proceedings of the 5th Symposium on Information Theory*. Moscow-Gorky, 1972, pp. 129—135.

[2] J. Shtarkov, "Coding of discrete sources with unknown statistics," in *Topics in Information Theory (Coll. Math. Soc. J. Bolyai, no. 16)*, I. Csiszár and P. Elias, Eds. Amsterdam, The Netherlands: North Holland, 1977, pp. 559–574.

[3] R. Krichevsky and V. Trofimov, "The preformance of universal coding," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199—207, March 1981.

[4] P. Laplace, *Philosphical essays on probabilities*, Translated by A. Dale from the 5th (1825) ed. Springer Verlag, New York, 1995.

[5] W. Gale and K. Church, "What is wrong with adding one?" in *Corpus based research into language*, N. Oostdijk and P. de Haan, Eds. Rodopi, Amsterdam, 1994, pp. 189—198.

[6] I. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3/4, pp. 237—264, December 1953.

[7] A. Orlitsky, N. Santhanam, and J. Zhang, "Always Good Turing: Asymptotically optimal probability estimation," in *Proceedings of the 44th Annual Symposium on Foundations of Computer Sciece*, October 2003.

[8] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*. Americam Mathematical Society, 2009.

[9] D. J. K. Farzan, "Coding schemes for chip-to-chip interconnect applications," *IEEE Transansactions on Very Large Scale Integration Systems*, vol. 14, no. 4, pp. 393–406, Apr 2006.

[10] F. Alajaji, "New results on the analysis of discrete communication channels with memory," Ph.D. dissertation, 1994.

[11] S. K. Gorantla and T. P. Coleman, "On reversible Markov chains and maximization of directed information," in *Proceedings of IEEE Symposium on Information Theory*, Seoul, South Korea, 2010, pp. 216–220.

[12] D. J. Aldous, "Random walks on finite groups and rapidly mixing Markov chains," in *Séminaire de Probabilités XVII - 1981/82, Springer Lecture Notes in Mathematics 986*, 1983.

[13] J. Rissanen, "A universal data compression system," *IEEE Transactions on Information theory*, vol. 29, no. 5, pp. 656–664, Sep 1983.

[14] P. Bühlmann and A. Wyner, "Variable length Markov chains," *Annals of Statistics*, vol. 27, no. 2, pp. 480–583, 1999.

[15] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via bic and mdl," *IEEE Transactions on Information theory*, vol. 52, no. 3, Mar 2006.

[16] A. Garivier, "Consistency of the unlimited BIC context tree estimator," *IEEE Transactions on Information theory*, vol. 52, no. 10, pp. 4630–4635, Sep 2006.

[17] I. Csiszár, "Large-scale typicality of Markov sample paths and consistency of MDL order estimators," *IEEE Transactions on Information theory*, vol. 48, no. 6, pp. 1616–1628, Jun 2002.

[18] I. Csiszár and P. C. Shields, "The consistency of the BIC Markov order estimator," *Annals of Statistics*, vol. 28, pp. 1601–1619, 2000.

[19] A. Galves, V. Maume-Deschamps, and B. Schmitt, "Exponential inequalities for VLMC empirical trees," *ESAIM: Probability and Statistics*, vol. 12, pp. 219–229, Jan 2008.

[20] A. Garivier and F. Leonardi, "Context tree selection: A unifying view," *Stochastic Processes and their Applications*, vol. 121, no. 11, pp. 2488–2506, Nov 2011.

[21] I. Csiszár and Z. Talata, "On rate of convergence of statistical estimation of stationary ergodic processes," *IEEE Transactions on Information theory*, vol. 56, no. 8, pp. 3637–3641, Aug 2010.

[22] A. Galves and F. Leonardi, "Exponential inequalities for empirical unbounded context trees," *In and Out of Equilibrium 2*, vol. 60, pp. 257–269, 2008.

[23] P. Vontobel, A. Kavcic, D. Arnold, and H. Loeliger, "A generalization of the blahut arimoto algorithm to finite state channels," *IEEE Transactions on Information Theory*, vol. 54, no. 5, May 2008.

[24] P. Sadeghi, P. Vontobel, and R. Shams, "Optimization of information rate upper and lower bounds for channels with memory," *IEEE Transactions on Information theory*, vol. 55, no. 2, pp. 663–688, Feb 2009.

[25] D. M. Arnold, H. Loeliger, P. Vontobel, A. Kavčić, and W. Zeng, "Simulation-based computation of information rates for channels with memory," *IEEE Transactions on Information theory*, vol. 52, no. 8, pp. 3498–3508, Aug 2006.

[26] J. Chen and P. H. Siegel, "Markov processes asymptotically achieve the capacity of Finite-State Intersymbol Interference channels," *IEEE Transactions on Information theory*, vol. 54, no. 3, pp. 1295–1303, March 2008.

[27] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: basic properties." *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[28] W. Feller, *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons; 2nd Edition, 1957, vol. 1.

[29] T. J. Tjalkens, Y. Shtarkov, and F. Willems, "Sequential weighting algorithms for multi-alphabet sources," in *6th Joint Swedish-Russian International Workshop on Information Theory*, August 1993, pp. 230–234.

[30] P. Ferrari and A. Galves, "Coupling and regeneration for stochastic processes," notes for a minicourse presented in XIII Escuela Venezolana de Matematicas.

[31] V. Guruswami, "Rapidly mixing Markov chains: A comparison of techniques," 2000, writeup at MIT Laboratory for Computer Science.

[32] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley and sons., 1991.